

UNIVERSITÀ
DI PAVIA

FACOLTÀ DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

TITOLO

Modelli di rappresentazione della conoscenza per il supporto alla decisione clinica in sistemi basati sull'Intelligenza Artificiale

Candidata: Sofia Buiatti

Relatore: Prof.ssa Lucia Sacchi
Correlatore: Prof. Enea Parimbelli

A.A. 2025/2026

Sommario

Abstract ... 3
Abstract (English version) ... 4
1 Introduzione ... 6
1.1 GUIDE-AI ... 6
1.2 Broncopneumopatia Cronica Ostruttiva (BPCO) ... 9
1.3 Linee Guida per la Pratica clinica e Sistemi Digitali di Supporto alla Decisione 13
1.4 Obiettivo della tesi ... 14
2 Intelligenza Artificiale nel supporto alla decisione clinica ... 17
2.1 Sistemi Esperti ... 17
2.1.1 Sistemi di supporto alla decisione tradizionali: il linguaggio PROforma ... 18
2.2 Natural Language Processing e Large Language Model ... 23
2.3 Retrieval-Augmented Generation (RAG) ... 30
2.3.1 Architettura RAG ... 31
2.3.2 Rappresentazione e recupero della conoscenza ... 32
2.3.3 L'utilizzo dei Large Language Models come motore ... 34
3 Rappresentazione della conoscenza ... 35
3.1 Ontologie ... 35
3.2 Knowledge Graph ... 36
3.3 Knowledge Graph per la rappresentazione di conoscenza medica ... 38
3.3.1 SNOMED CT ... 38
3.3.2 ATC ... 39
3.3.3 UMLS ... 40
3.4 Vector DB ... 40
4 Estrazione delle entità nel testo ... 43
4.1 Architettura del Concept Extractor ... 44
4.2 Approccio basato su dizionario ... 45
4.2.1 Corrispondenza lessicale esatta ... 46
4.2.2 Corrispondenza lessicale approssimata ... 47
4.3 Approccio neurale ... 48
4.3.1 Riconoscimento delle entità (NER) ... 48
4.3.2 Rappresentazione vettoriale e recupero ... 50

4.4 Disambiguazione e riordinamento dei candidati 52
4.5 Metriche di valutazione 54
5 Implementazione e Risultati 56
5.1 Pipeline 56
5.2 Flowchart 57
5.3 Modulo strutturato compilabile 65
5.4 Formalizzazione della linea guida tramite linguaggio PROforma 68
5.5 Knowledge Graph 80
5.5.1 Base della conoscenza 81
5.5.2 Modellazione delle Linee Guida nel database a grafo 87
5.5.3 Integrazione con SNOMED CT e ATC 91
5.6 Estrattore dei concetti SNOMED CT 97
5.6.1 Costruzione dei Dizionari 97
5.6.2 Creazione del file di validazione 98
5.6.3 Performance dell'estrattore di concetti 110
6 Conclusioni e sviluppi futuri 115
6.1 Conclusioni 115
6.2 Sviluppi futuri 118
Appendice 1 – Prompt per traduzione delle figure in formato XML 120
Appendice 2 – Dati riguardanti la creazione della ground truth 121
Appendice 3 – Concetti estratti e validati d'esempio 124
Concetti estratti dell'esempio al paragrafo 5.6.3 124
Ground truth dell'esempio al paragrafo 5.6.3 125
Bibliografia 126

Abstract

Il presente lavoro di tesi si colloca all'interno del progetto GUIDE-AI, finanziato dalla Commissione Europea. L'obiettivo di GUIDE-AI è di sfruttare l'intelligenza artificiale per aumentare l'aderenza alle linee guida cliniche nella gestione dei percorsi terapeutici di pazienti affetti da malattie croniche in Europa.

Un numero elevato di pazienti non riceve la terapia farmacologica ottimale raccomandata, anche a causa della lunghezza e della complessità delle linee guida. Il progetto si pone l'obiettivo di incrementare l'aderenza a percorsi terapeutici fondati su evidenze scientifiche e, di conseguenza, di ridurre il tasso di mortalità attribuibile a malattie croniche in Europa.

GUIDE-AI viene sviluppato come modello di riferimento per quattro indicazioni cliniche: broncopneumopatia cronica ostruttiva, scompenso cardiaco cronico, malattia renale cronica e asma.

Il presente lavoro di tesi si inserisce all'interno delle fasi iniziali del progetto GUIDE-AI e si focalizza sull'analisi della linea guida clinica per la broncopneumopatia cronica ostruttiva (BPCO o Chronic obstructive pulmonary disease, COPD), approfondendone l'intero percorso terapeutico. L'attività si articola in quattro macro-fasi sequenziali. La prima fase è la formalizzazione della conoscenza clinica contenuta nella linea guida *Global Initiative for Chronic Obstructive Lung Disease* (GOLD, report 2026), tradotta in un flowchart strutturato che mappa l'intero percorso terapeutico, dalla presa in carico iniziale al follow-up. La seconda fase ha previsto la rappresentazione formale di tale conoscenza all'interno di un sistema esperto utilizzando il linguaggio PROforma per mezzo del software Composer, garantendo un comportamento deterministico, verificabile e privo di allucinazioni. La terza fase prevede la costruzione della base di conoscenza su cui poi si baserà l'architettura di *Retrieval Augmented Generation* (RAG) del progetto, implementando un grafo di conoscenza (*Knowledge Graph*) con le linee guida di riferimento. La quarta fase prevede l'utilizzo di una pipeline

per l'estrazione automatica di concetti clinici e la loro mappatura alla nomenclatura standardizzata SNOMED CT. Quest'ultimo passaggio è essenziale per completare l'implementazione del grafo di conoscenza, rendendo possibile il legame tra i capitoli della linea guida e i concetti standardizzati SNOMED CT.

Nel complesso, il lavoro dimostra l'importanza di utilizzare sistemi di intelligenza artificiale basati sulla conoscenza in contesti clinici e la loro fattibilità tramite sistemi esperti e RAG.

Abstract (English version)

This thesis is part of the GUIDE-AI project, funded by the European Commission. The aim of GUIDE-AI is to leverage artificial intelligence to increase adherence to clinical guidelines in the management of therapeutic pathways for patients with chronic diseases in Europe.

Because many patients do not receive the recommended optimal pharmacological therapy, owing to the length and complexity of the guidelines, the project aims to increase adherence to evidence-based therapeutic pathways and thereby reduce the mortality rate attributable to chronic diseases in Europe. GUIDE-AI is being developed as a reference model for four clinical indications: chronic obstructive pulmonary disease, chronic heart failure, chronic kidney disease and asthma.

This thesis is situated in the early phases of the GUIDE-AI project and focuses on the analysis of the clinical guideline for chronic obstructive pulmonary disease (COPD), examining the entire therapeutic pathway. The work is organized into four sequential macro-phases. The first phase concerned the formalization of the clinical knowledge contained in the Global Initiative for Chronic Obstructive Lung Disease (GOLD, 2026 report) guideline, translated into a structured flowchart that maps the entire therapeutic pathway, from initial patient intake to follow-up. The second phase involved the formal representation of this knowledge within an expert system, using the PROforma language by means of the Composer software, ensuring deterministic, verifiable behavior that is free of hallucinations. The third

phase consists of building the knowledge base on which the project's Retrieval-Augmented Generation (RAG) architecture will subsequently rely, by implementing a Knowledge Graph from the reference guideline. The fourth phase consists of using a pipeline for the automatic extraction of clinical concepts and their mapping to the standardized SNOMED CT nomenclature. This last step is essential to complete the implementation of the Knowledge Graph, as it makes it possible to link the chapters of the guideline to standardized SNOMED CT concepts.

Overall, the work demonstrates the importance of using knowledge-based artificial intelligence systems in clinical settings, as well as their feasibility through expert systems and RAG.

1 Introduzione

1.1 GUIDE-AI

Il presente lavoro di tesi si colloca all'interno del progetto GUIDE-AI, finanziato dalla Commissione Europea nell'ambito del programma Innovative Health Initiatives (GA n. 101253015).[1] L'obiettivo di GUIDE-AI è di sfruttare l'intelligenza artificiale (IA) per aumentare l'aderenza alle linee guida durante la gestione dei percorsi terapeutici di pazienti affetti da malattie croniche in Europa.

Le malattie croniche polmonari, cardiache e renali rappresentano una quota rilevante delle patologie croniche nella popolazione europea, con quasi l'85% dei decessi nell'UE ad esse attribuibili.[2] La terapia farmacologica basata sulle linee guida è fondata su evidenze scientifiche e costituisce il punto di riferimento del trattamento farmacologico per pazienti affetti da malattie croniche. Tuttavia, un numero elevato di pazienti non riceve la terapia farmacologica ottimale raccomandata, poiché spesso la patologia viene gestita da medici non specialisti o non aggiornati. I report delle linee guida, infatti, vengono modificati alla luce di nuove evidenze scientifiche e questo rende più difficile l'allineamento dei professionisti.

Il progetto GUIDE-AI si pone tre obiettivi principali:

- Centralità della persona: GUIDE-AI integra le prospettive di pazienti, medici, amministratori sanitari ed enti finanziatori, affrontando sia l'assistenza ospedaliera che ambulatoriale e coprendo l'intero continuum terapeutico, dall'inizio della terapia al monitoraggio a lungo termine e agli aggiustamenti del trattamento.
- Innovazione:
    1. sviluppare software di IA generativa, basati su modelli linguistici (Large Language Models, LLM), in grado di sfruttare le linee guida mediche per supportare il medico nella definizione del percorso terapeutico;

2. progettare l’infrastruttura tecnica che permetta l’interfaccia con l’utente e la ricezione di dati clinici;
3. verificare il funzionamento dell’IA tramite dataset retrospettivi, ovvero cartelle cliniche storiche. Questo permette di misurare le prestazioni del software, confrontando la terapia suggerita dall’IA con quella proposta dai medici e con quella che la linea guida avrebbe raccomandato.

- Valutazione in ambiente reale: GUIDE-AI verrà valutato mediante uno studio clinico in ambiente reale basato su un disegno randomizzato a due bracci. Verranno perciò suddivisi i medici partecipanti in due gruppi: il braccio di controllo, in cui i medici cureranno i pazienti senza l’ausilio dell’IA, e il braccio sperimentale, in cui i medici saranno supportati da GUIDE-AI. Al termine dello studio, si confronteranno i risultati dei due gruppi per verificare se l’IA abbia portato ad un miglioramento della percentuale di terapie conformi alle linee guida.

GUIDE-AI riunisce partner privati ed istituzioni accademiche di spicco al fine di superare le soluzioni esistenti e incrementare la terapia medica basata sulle linee guida, impiegando l’IA generativa e rendendo applicabili le più recenti evidenze scientifiche.[2]

Il sistema GUIDE-AI, mostrato in Figura 1, si articola in un’architettura modulare organizzata in cinque fasi, progettata per supportare il processo decisionale clinico nella gestione terapeutica di pazienti cronici.

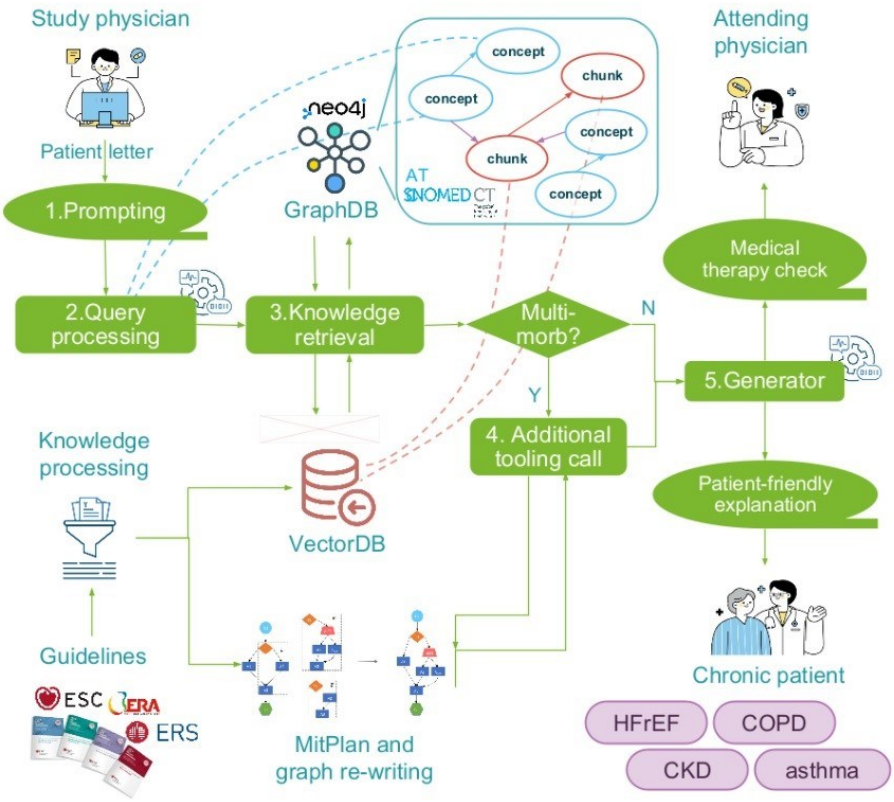


Figura 1: architettura di GUIDE-AI.

Il processo ha inizio con la fase di *prompting* (1), durante la quale il medico responsabile dello studio inserisce i dati clinici del paziente come input al sistema (richiesta o query). Questa richiesta viene sottoposta a una fase di elaborazione (2), finalizzata alla formulazione di un'interrogazione strutturata che porti al recupero delle informazioni rilevanti.

Affinché l'interrogazione sia possibile, un modulo dedicato all'elaborazione della conoscenza processa le linee guida cliniche di riferimento e costruisce la base di conoscenza. Questa viene strutturata in due rappresentazioni complementari e sinergiche.

La prima è un database a grafo, implementato in Neo4j, che modella le relazioni semantiche tra concetti clinici codificati secondo le terminologie *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT) e *Anatomical Therapeutic Chemical Classification System* (ATC), e segmenti testuali (chunk).

La seconda è un database vettoriale, che consente il recupero dei *chunk* per similarità semantica.

La fase di recupero della conoscenza (3) interroga entrambe le sorgenti, database a grafo e vettoriale, per identificare gli elementi clinici pertinenti al caso in esame e procedere alla generazione della risposta.

Nella fase conclusiva di generazione, il sistema produce due output distinti: un riscontro di verifica della terapia farmacologica, destinato al medico curante ai fini della validazione clinica, e una spiegazione in lessico più accessibile, destinata alla comunicazione col paziente.

GUIDE-AI viene sviluppato basandosi su quattro patologie croniche: broncopneumopatia cronica ostruttiva (BPCO), asma, scompenso cardiaco cronico (SCC) e malattia renale cronica (MRC). Questa configurazione potrà essere estesa ad altre patologie croniche.

Il contributo specifico del presente lavoro di tesi si inserisce all'interno delle fasi iniziali del progetto GUIDE-AI e si focalizza sull'analisi delle linee guida cliniche per la BPCO, approfondendone l'intero percorso terapeutico.

1.2 Broncopneumopatia Cronica Ostruttiva (BPCO)

La broncopneumopatia cronica ostruttiva (BPCO) è una patologia polmonare disomogenea; essa colpisce i polmoni in modo disuguale, variando in gravità e aspetto tra diverse regioni, lobi o compartimenti tessutali. La BPCO si manifesta con sintomi respiratori cronici (dispnea, tosse, produzione di espettorato ed esacerbazioni o riacutizzazioni), correlati ad ostruzioni persistenti e, spesso, progressive alle vie aeree. [3], [4]

La BPCO deriva da interazioni gene-ambiente che si verificano durante la vita dell'individuo e che possono danneggiare i polmoni, alterare il loro sviluppo e accelerarne l'invecchiamento cellulare.

Il fumo di tabacco (assunto tramite sigaretta, pipa o sigaro) e il consumo di marijuana rappresentano i principali fattori di rischio ambientali per la BPCO nei paesi ad alto reddito, dove coprono oltre il 70% dei casi. [5] L'esposizione passiva e cronica al fumo di tabacco ambientale contribuisce alla comparsa di sintomi respiratori. [6]

Nei paesi a basso e medio reddito (LMIC), che da soli contribuiscono ad oltre l'85% del carico mondiale della malattia, il fumo incide solo per il 30%-40%.[7]

L'inquinamento atmosferico, composto da particolato fine, ozono, ossidi di azoto o di zolfo, metalli pesanti e gas serra, [6], [7] è responsabile di circa il 50% del rischio di BPCO nei LMIC. [9]

Il fattore di rischio genetico più rilevante per la BPCO è rappresentato dalle mutazioni nel gene SERPINA1 che portano al deficit ereditario di α-1 antitripsina (AATD). [10], [11]

In un contesto clinico in cui i sintomi e i fattori di rischio sono riconducibili alla BPCO, la presenza di un'ostruzione persistente e non completamente reversibile del flusso aereo (FEV1/FVC < 0,7 post-broncodilatazione), misurata tramite spirometria, conferma la diagnosi.

Il percorso terapeutico della BPCO si basa, secondo le linee guida della *Global Initiative for Chronic Obstructive Lung Disease* (GOLD)[3], sull'entità dei sintomi e sul rischio di riacutizzazioni. La classificazione GOLD identifica tre gruppi di rischio, definiti sulla base del numero di pregresse esacerbazioni moderate e severe e dei punteggi di sintomatologia riportati dal paziente tramite questionari come *modified Medical Research Council dyspnea scale* (mMRC) o *COPD Assessment Test* (CAT).

La scala mMRC è stata il primo questionario sviluppato per misurare la dispnea, che è un sintomo chiave in molti pazienti affetti da BPCO.[12]

Il questionario chiede di selezionare il grado che descrive meglio il proprio stato: [12]

- mMRC Grado 0: Ho affanno solo per sforzi intensi.
- mMRC Grado 1: Mi viene l'affanno se cammino in fretta in pianura o su una leggera salita.
- mMRC Grado 2: In pianura cammino più lentamente dei coetanei a causa dell'affanno, oppure devo fermarmi per respirare quando cammino al mio passo.
- mMRC Grado 3: Devo fermarmi per respirare dopo aver camminato in pianura per circa 100 metri o dopo pochi minuti.
- mMRC Grado 4: Ho troppo affanno per uscire di casa, oppure ho affanno anche quando mi vesto o mi spoglio.

Il CAT è un questionario composto da 8 domande che valuta lo stato di salute nei pazienti con BPCO. È stato sviluppato per essere applicabile in tutto il mondo e sono disponibili traduzioni validate in un'ampia gamma di lingue. Il punteggio varia da 0 a 40. [13], [14]

Il questionario chiede di indicare un punteggio da 0 a 5 per ogni voce: [13]

- Tosse: 0 indica “Non tossisco mai” e 5 indica “Tossisco continuamente”
- Espettorato (muco/catarro): 0 indica “Non ho affatto catarro (muco) nel petto” e 5 indica “Ho il petto completamente pieno di catarro (muco)”
- Oppressione toracica: 0 indica “Non sento alcuna oppressione al petto” e 5 indica “Sento una forte oppressione al petto”
- Dispnea (affanno): 0 indica “Quando cammino in salita o salgo una rampa di scale non ho affanno” e 5 indica “Quando cammino in salita o salgo una rampa di scale ho molto affanno”
- Limitazioni a casa: 0 indica “Non sono limitato in nessuna attività a casa” e 5 indica “Sono molto limitato nelle attività a casa”

- Sicurezza a uscire: 0 indica “Mi sento sicuro a uscire di casa nonostante la mia condizione polmonare” e 5 indica “Non mi sento affatto sicuro a uscire di casa a causa della mia condizione polmonare”
- Sonno: 0 indica “Dormo profondamente” e 5 indica “Non dormo profondamente a causa della mia condizione polmonare”
- Energia: 0 indica “Ho moltissima energia” e 5 indica “Non ho alcuna energia”

Al termine del questionario si sommano le voci per calcolare il punteggio totale.

Una volta ottenuti tali punteggi è possibile classificare la patologia del paziente: [3]

1. gruppo A: pochi sintomi (mMRC≤ 1 o CAT <10) e basso rischio di esacerbazione (riacutizzazioni=0);
2. gruppo B: molti sintomi (mMRC≥ 2 o CAT ≥ 10) e basso rischio di esacerbazione (riacutizzazioni=0);
3. gruppo E: alto rischio di esacerbazione, indipendentemente dai sintomi (riacutizzazioni≥ 1).

Il cardine della terapia farmacologica di mantenimento è rappresentato dai broncodilatatori a lunga durata d’azione, distinti in due classi principali: i beta2-agonisti a lunga durata d’azione (*long-acting β2-agonist*, LABA) e gli antimuscarinici a lunga durata d’azione (*long-acting muscarinic antagonists*, LAMA). Questi farmaci possono essere somministrati in monoterapia o, più frequentemente, in combinazione (LABA+LAMA), con l’obiettivo di massimizzare la broncodilatazione e migliorare il volume espiratorio massimo nel primo secondo (*Forced Expiratory Volume in the first second*, FEV1). [15]

Nei pazienti appartenenti al gruppo E, la scelta terapeutica iniziale tiene conto anche del tasso di eosinofili nel sangue, usato come biomarcatore di infiammazione. Nel caso in cui il valore degli eosinofili sia maggiore di 300

cellule/μL è indicato combinare alla duplice terapia broncodilatatoria un trattamento con corticosteroidi inalatori (*inhaled corticosteroids*, ICS).[16], [17]

Trattandosi di una patologia cronica e progressiva, la gestione della BPCO prevede controlli periodici, durante i quali il medico valuta l'eventuale necessità di intensificare la terapia in caso di persistenza o peggioramento della sintomatologia e delle riacutizzazioni. In caso di esacerbazioni persistenti in corso di monoterapia broncodilatatoria, è raccomandata l'escalation a LABA+LAMA; qualora le esacerbazioni continuino nonostante la duplice terapia, si suggerisce l'ulteriore escalation a LABA+LAMA+ICS. [18]

I pazienti con BPCO presentano frequentemente altre malattie concomitanti (comorbidità) che influenzano anch'esse la condizione clinica e la prognosi, e che richiedono un trattamento specifico.

1.3 Linee Guida per la Pratica clinica e Sistemi Digitali di Supporto alla Decisione

Le Linee Guida per la Pratica Clinica sono state definite nel 2011 dall'Institute of Medicine (IOM) come "documenti che includono raccomandazioni finalizzate a ottimizzare l'assistenza al paziente fondate su una revisione sistematica delle prove di efficacia e su una valutazione di benefici e danni di opzioni assistenziali alternative". [19]

Il ruolo delle linee guida in Italia è riportato anche nel contesto normativo; infatti, nell'articolo 5 della legge n. 24/2017 sulla responsabilità professionale e le buone pratiche clinico-assistenziali, si legge che "gli esercenti delle professioni sanitarie, nell'esecuzione delle prestazioni sanitarie con finalità preventive, diagnostiche, terapeutiche, palliative, riabilitative e di medicina legale, si attengono, salve le specificità del caso concreto, alle raccomandazioni previste dalle linee guida […] elaborate da enti e istituzioni pubblici e privati nonché dalle società scientifiche e dalle associazioni tecnico-scientifiche delle professioni sanitarie iscritte in

apposito elenco istituito e regolamentato con decreto del Ministro della salute" [20].

Sebbene le linee guida rappresentino un riferimento fondamentale per la pratica clinica e il loro rispetto assuma rilevanza anche sul piano normativo, la loro effettiva integrazione nei processi decisionali può presentare alcune criticità. La complessità dei documenti, la quantità di informazioni da analizzare e la necessità di individuare rapidamente le raccomandazioni pertinenti al singolo caso possono rendere difficile una consultazione sistematica durante l'attività assistenziale. A ciò si aggiunge la necessità per i professionisti sanitari di mantenersi costantemente aggiornati sulle revisioni delle raccomandazioni, che possono essere modificate alla luce di nuove evidenze scientifiche.

Per questi motivi, negli anni, uno dei tanti obiettivi della bioingegneria è stato quello di sviluppare servizi di supporto alle decisioni cliniche che si basassero sulla formalizzazione delle linee guida di riferimento in modo da renderle interpretabili da un computer e facilmente utilizzabili dai clinici attraverso dei supporti digitali.

Questi servizi rientrano tra i cosiddetti sistemi di supporto alle decisioni cliniche (in inglese *Clinical Decision Support Systems*, CDSS) basati su conoscenza. Si tratta di strumenti progettati per assistere i medici e gli operatori sanitari nei processi decisionali clinici legati alla diagnosi, alla terapia e, più in generale, alla gestione dei pazienti, utilizzando una base di conoscenza esplicita, che nel caso specifico è costituita dalle linee guida cliniche.

I CDSS formalizzati basati su linee guida informatizzate offrono un supporto rapido e accurato nel processo decisionale clinico, superando le limitazioni legate alla consultazione manuale dei documenti.

1.4 Obiettivo della tesi

Il progetto GUIDE-AI si pone l'obiettivo di sviluppare un sistema IA per il supporto alla decisione clinica per i percorsi terapeutici di malattie croniche, che faccia riferimento ad una base di conoscenza costituita dalle linee guida cliniche.

In particolare, il presente lavoro si concentra sulla linea guida relativa alla BPCO e si propone di studiare, analizzare e formalizzare la conoscenza clinica in essa contenuta. A tal fine, è stato inizialmente realizzato un diagramma a blocchi volto a rappresentare il percorso terapeutico di mantenimento descritto nella linea guida.

Ciò ha permesso di identificare le variabili cliniche rilevanti per generare un suggerimento terapeutico coerente con le raccomandazioni di riferimento.

Successivamente, la conoscenza clinica è stata modellizzata attraverso due approcci. Il primo è rappresentato dai sistemi esperti tradizionali, basati sulla rappresentazione esplicita delle conoscenze e delle regole decisionali. Il secondo è costituito da un'architettura *Retrieval-Augmented Generation* (RAG), che combina il recupero di informazioni da una base di conoscenza con la generazione di risposte mediante un modello linguistico. La tesi si propone di descrivere e confrontare i due approcci, analizzandone le caratteristiche, i punti di forza e i limiti.

Nell'ambito del progetto GUIDE-AI, l'architettura RAG è stata individuata come approccio per l'implementazione del sistema. Per supportarne il funzionamento, una fase centrale del lavoro è stata dedicata alla costruzione di un grafo di conoscenza contenente le informazioni relative alla farmacoterapia di mantenimento della BPCO, estratte dalla linea guida di riferimento e organizzate in una rappresentazione strutturata.

Per favorire l'integrazione di tali informazioni con una terminologia clinica standardizzata, è stato inoltre previsto il collegamento del grafo di conoscenza con quello relativo a SNOMED CT. A questo scopo, è stato utilizzato e adattato un estrattore di concetti sviluppato nell'ambito di un progetto precedente e affine, con l'obiettivo di automatizzare il processo di identificazione e associazione dei concetti clinici.

Nel complesso, il lavoro mira a esplorare diverse modalità di rappresentazione e utilizzo della conoscenza clinica, ponendo le basi per un sistema di supporto alle

decisioni cliniche in grado di recuperare e utilizzare le raccomandazioni contenute nelle linee guida in modo strutturato e contestualizzato. Il confronto tra i sistemi esperti e l'architettura RAG, insieme alla costruzione del grafo di conoscenza e alla sua integrazione con SNOMED CT, costituisce pertanto il contributo principale della tesi allo sviluppo del progetto GUIDE-AI.

2 Intelligenza Artificiale nel supporto alla decisione clinica

2.1 Sistemi Esperti

Un Sistema Esperto è un programma informatico che: [21]

- ragiona utilizzando conoscenze specifiche del dominio, sia di natura simbolica che matematica;
- impiega metodi specifici del dominio, sia euristici (basati sulla plausibilità) che algoritmici (basati sulla certezza);
- opera con prestazioni paragonabili a quelle degli specialisti nel proprio ambito di riferimento;
- rende comprensibili sia le proprie conoscenze sia le motivazioni alla base delle risposte fornite.

I Sistemi Esperti sono diventati una componente dell'intelligenza artificiale tra la fine degli anni '60 e l'inizio degli anni '70 e sono sistemi di supporto alla decisione che si basano sulla rappresentazione esplicita e formale della conoscenza (*Knowledge-Based*). [21]

Un sistema esperto si articola in tre componenti:

1. *Knowledge Base* (KB) è il nucleo centrale del sistema, all'interno del quale viene memorizzato il sapere strutturato del dominio di riferimento. Contiene fatti assiomatici e regole logiche che governano il dominio;
2. il motore di inferenza (*reasoning engine*) rappresenta l'unità di calcolo e ragionamento del sistema. È un algoritmo formale che analizza i dati clinici alla luce del contenuto della base di conoscenza per formulare soluzioni e raccomandazioni;
3. l'interfaccia utente consente di interagire col sistema, ricevendo in input i dati e fornendo spiegazioni circa la soluzione proposta e il ragionamento sviluppato dal sistema.

In ambito medico, la conoscenza è strutturata all'interno delle linee guida per la pratica clinica. Esse si propongono di uniformare il trattamento di pazienti affetti da patologie simili in ambienti sanitari differenti, ottimizzare l'efficacia delle cure e diffondere una rappresentazione di processi complessi che richiedono competenze differenti.

I sistemi esperti combinano la conoscenza contenuta nelle linee guida informatizzate con i dati clinici dei pazienti per fornire raccomandazioni paziente-specifiche.

Sistemi di supporto alle decisioni cliniche basati sulla formalizzazione delle linee guida sono un approccio che presenta i seguenti vantaggi:

- l'intero algoritmo decisionale è trasparente e spiegabile, il sistema è in grado di esplicitare l'esatto percorso logico e le specifiche regole cliniche che hanno condotto a una determinata raccomandazione terapeutica;
- operando dentro confini di un'architettura chiusa e deterministica, non vi possono essere allucinazioni. Si garantisce così la massima aderenza alle direttive imposte.

2.1.1 Sistemi di supporto alla decisione tradizionali: il linguaggio PROforma

Tra gli strumenti per lo sviluppo dei CDSS tradizionali basati su conoscenza si distingue PROforma: un linguaggio formale di rappresentazione della conoscenza sviluppato nel 1996 dall'Advanced Computation Laboratory of Cancer Research UK[22] e ora distribuito da Deontics Ltd[23]. PROforma è un linguaggio finalizzato alla modellizzazione dei processi clinici e alla formalizzazione delle linee guida, al fine di supportare il processo decisionale e la gestione del workflow clinico. Il linguaggio PROforma è un formato di rappresentazione delle linee guida di tipo "task-based", che consiste nella definizione di regole formali contestualizzate all'interno di un processo clinico rappresentato come una serie di attività (o task). Una regola, ad esempio, può

rappresentare una precondizione per uno specifico intervento clinico oppure una controindicazione a una particolare terapia farmacologica. PROforma combina le caratteristiche di un linguaggio di specifica formale, sviluppato nell'ambito dell'ingegneria del software, con quelle dei linguaggi di rappresentazione della conoscenza sviluppati nell'ambito dell'intelligenza artificiale. [24]

La struttura del linguaggio PROforma è basata su un modello semplice ma versatile chiamato "modello domino", rappresentato in Figura 2. Questo descrive le relazioni e i processi che legano le azioni, le opzioni, le decisioni, i piani, gli obiettivi e le soluzioni candidate: dato un certo insieme di convinzioni, un agente può definire degli obiettivi e diverse soluzioni per questi ultimi. Se ci sono più opzioni, ad esempio diagnosi o trattamenti alternativi, l'agente deve considerare gli argomenti a favore e quelli contro ciascuna di esse, quindi prendere decisioni sulla base di questi argomenti.

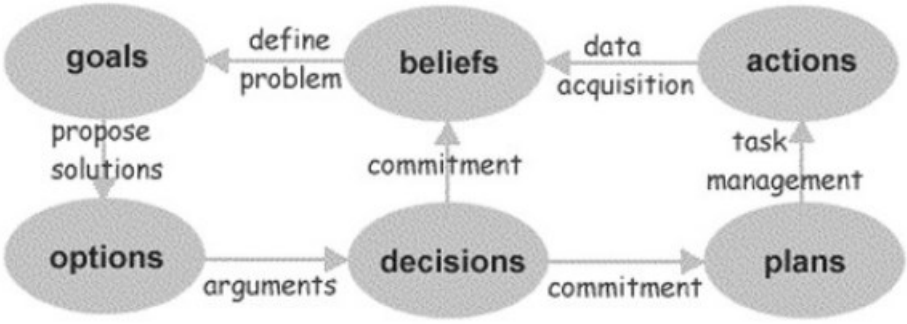


Figura 2: rappresentazione del Modello Domino. I quattro elementi di sinistra del diagramma costituiscono la parte relativa alle decisioni, quelli di destra il piano e le relative azioni da mettere in atto [24].

La nascita di PROforma deriva da un processo tramite il quale il modello sopra citato è stato concretizzato in un insieme di task. I task sono oggetti software formali che possono essere composti per formare delle reti che rappresentano linee guida cliniche o altri processi. La formalizzazione con questo linguaggio prevede l'utilizzo di un set di data item e quattro tipologie di task, la cui ontologia è mostrata in Figura 3:

- Action: permette di definire un'azione da consigliare direttamente all'utente, senza che quest'ultimo possa effettuare una scelta o interagire con la decisione. Una action mostra un messaggio da leggere, che indica l'azione che dovrebbe essere eseguita a quel punto del processo clinico (e.g. somministrare un determinato farmaco).
- Enquiry: rappresenta il punto di una linea guida in cui le informazioni devono essere acquisite, e quindi memorizzate, per essere fruibili dal motore inferenziale (e.g. raccolta dei sintomi del paziente).
- Decision: permette all'utente di scegliere tra una serie di opzioni, chiamate "candidates". L'obiettivo di questo task è quello di far valutare all'utente una o più soluzioni candidate, raccomandandone alcune e sconsigliandone altre in base a specifiche regole che tengono conto delle evidenze a disposizione (e.g. scegliere quale farmaco somministrare sulla base dei sintomi del paziente).
- Plan: permette di raggruppare diversi task che sono logicamente o temporalmente legati tra loro, e di comporli per poter eseguire operazioni complesse.

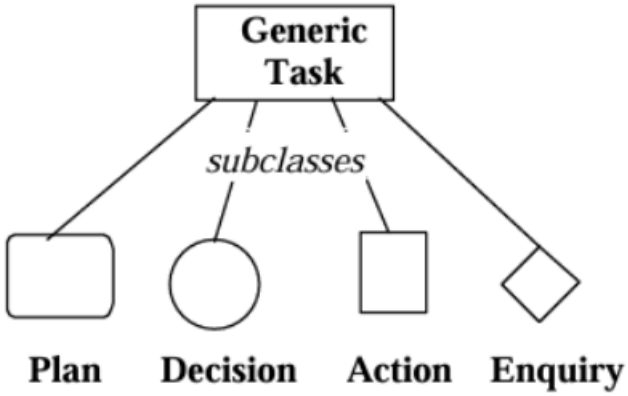


Figura 3: ontologia dei task di PROforma [24].

Essendo sottoclassi di un generico task, Plan, Decision, Action ed Enquiry hanno delle proprietà comuni, tra cui, ad esempio[24]:

- Description: spiegazione del task e rimando a documenti esterni che giustifichino e spieghino le operazioni consigliate;
- Precondition: condizioni logiche che devono essere vere affinché il task abbia inizio;
- Postcondition: condizioni logiche che devono esser vere quando il task termina.

In aggiunta a quelle comuni, ogni tipologia di task ha delle proprietà specifiche descritte in Tabella 1 [24].

Tipi di task	Proprietà
Decision	**Candidates**: opzioni tra cui scegliere e a cui sono associate delle regole chiamate arguments che le consigliano o sconsigliano sulla base dei dati raccolti.
	Choice mode: permette di definire se la decisione può prevedere la selezione di un solo candidato (single) o di più candidati (multiple).
Plan	**Abort condition**: condizioni logiche che, se vere, fanno terminare il Plan senza successo.
	Termination Condition: condizioni logiche che, se vere, fanno terminare il Plan anche se alcuni dei suoi task non sono stati ancora eseguiti.
Enquiry	**Sources**: set di data item per i quali viene richiesto di associare un valore.
Action	**Procedure**: testo libero tramite cui si definisce l'azione che deve compiere l'utente.

Tabella 1: proprietà specifiche dei singoli task.

Per la formalizzazione delle linee guida con PROforma, Deontics rende disponibili due applicazioni. Il Composer consente di creare, modificare e visualizzare graficamente i task che formalizzano le linee guida. Il Tester è un ambiente web-based che permette al programmatore di testarle e farne il debug, richiamando il motore [25].

A partire dai primi anni 2000, la nascita di questo linguaggio ha favorito lo sviluppo di nuove tecnologie e applicazioni nell'assistenza sanitaria, sia primaria che specialistica. Tali applicazioni sono state impiegate in diversi ambiti: il supporto alla prescrizione di farmaci, la valutazione del rischio nell'assistenza primaria, la valutazione e il trattamento di malattie croniche e l'assistenza ospedaliera[24].

Queste applicazioni sono state valutate e hanno dimostrato come il supporto decisionale basato su linee guida formalizzate diminuisca notevolmente il rischio che i clinici facciano delle scelte che deviano dalle linee guida e quindi dalle migliori evidenze mediche disponibili. [26]

Tuttavia, la formalizzazione con il linguaggio PROforma è un'operazione non banale che richiede molto tempo. Ciò rappresenta un limite rilevante soprattutto negli ambiti della medicina in cui la ricerca scientifica progredisce rapidamente e le linee guida subiscono continui aggiornamenti. In questi contesti, risulta difficile mantenere tale approccio costantemente aggiornato. Inoltre, i sistemi esperti hanno una struttura rigida che non permette di gestire scenari clinici non previsti. Questa rappresentazione della conoscenza impone che ogni scenario che si vuole rappresentare debba essere modellizzato, dal momento che il motore d'inferenza si basa unicamente sulla base di conoscenza fornita.

2.2 Natural Language Processing e Large Language Model

L'elaborazione del linguaggio naturale (*Natural Language Processing*, NLP) è una branca dell'Intelligenza Artificiale che si occupa di sviluppare sistemi in grado di riconoscere, comprendere e generare il linguaggio naturale. Questo paradigma combina algoritmi di *machine learning* e *deep learning* e si basa su diversi processi e tecniche che consentono una corretta interpretazione del linguaggio umano, quali: [27]

- Tokenizzazione: suddivisione del testo in unità base chiamate token;
- Parsing: analisi della struttura sintattica delle frasi;

- Lemmatizzazione e *stemming*: operazioni alternative di normalizzazione morfologica. La lemmatizzazione riduce una parola alla sua forma base (lemma), mentre lo *stemming* ne individua la radice eliminando i suffissi.
- Riconoscimento delle parti del discorso (*part of speech tagging,* POS): assegnazione a ciascuna parola di una categoria grammaticale (nome, verbo, aggettivo)
- Analisi semantica: assegnazione di un significato alle parole
- Analisi del contesto: interpretazione del significato delle parole, valutandolo all'interno di un quadro complessivo.

Le principali applicazioni dell'NLP sono molteplici[27]:

- Estrazione di Informazioni: processo cardine per la conversione di testi non strutturati in rappresentazioni formalmente organizzate dell'informazione, mediante l'identificazione e la classificazione di entità semanticamente rilevanti. All'interno di questo quadro metodologico si collocano task specifici quali la *Named Entity Recognition* (NER), finalizzata all'individuazione e classificazione delle entità rilevanti menzionate nel testo[28], e l'*Entity Linking*, volto alla disambiguazione e all'associazione delle entità individuate a basi di conoscenza di riferimento.[29]

- Sistemi di Domanda e Risposta (Question Answering, QA): architetture computazionali progettate per l'estrazione e la generazione automatica di risposte precise a quesiti espressi in linguaggio naturale. Tali sistemi si articolano principalmente in due categorie. I modelli basati sulla conoscenza sono focalizzati sul recupero dell'esatta porzione di testo contenente la risposta all'interno di un corpus documentale. I modelli generativi sono capaci di sintetizzare risposte articolate attingendo alla conoscenza parametrica dei modelli o integrando basi di conoscenza esterne.[30]

- Categorizzazione e Modellazione dei Temi: algoritmi di clustering e classificazione volti all'indicizzazione semantica di documenti, al rilevamento di duplicati informativi e alla generazione di alert predittivi su contenuti rilevanti.
- Analisi dei Sentimenti: identificazione computazionale del tono emotivo, dei giudizi soggettivi o dello stato psicologico espresso all'interno di testi, utile per l'analisi delle opinioni e il monitoraggio dei feedback.
- Sintesi e Traduzione Automatica: riassunto di testi estesi e conversione multilingua.
- Speech-to-Text & Text-to-Speech: sistemi di trascrizione automatica della dettatura e di lettura di testi.

L'evoluzione storica dell'NLP ha seguito un percorso graduale a partire dagli anni '50 sino ad oggi. I primi tentativi di traduzione automatica si basavano su regole linguistiche codificate manualmente[31], mentre nel 1966 il programma ELIZA dimostrò la fattibilità dei primi sistemi di dialogo, simulando conversazioni psicoterapeutiche tramite pattern di frasi predefiniti.[32] Negli anni '90, con la crescente disponibilità di dati digitali, l'NLP ha iniziato a sfruttare approcci statistici capaci di stimare sequenze linguistiche in termini di probabilità.[33] A partire dagli anni 2000, lo sviluppo di algoritmi di *machine learning* supervisionato ha permesso un significativo avanzamento delle applicazioni di elaborazione del linguaggio naturale, superando i precedenti approcci basati su regole. [34]Successivamente, l'introduzione delle reti neurali profonde ha consentito di gestire con maggiore efficacia sequenze testuali articolate, gettando le basi per l'evoluzione verso i moderni modelli Transformer. [35]

L'introduzione dell'architettura Transformer, formalizzata da Vaswani et al. nel 2017 [35], ha ridefinito lo stato dell'arte dell'NLP. I modelli tradizionali elaboravano i testi in modo strettamente sequenziale, una parola alla volta; questo

approccio rendeva difficile la contestualizzazione semantica all'interno di testi complessi.

L'architettura Transformer risolve tali criticità attraverso il meccanismo di *self-attention*. Questo operatore matematico valuta dinamicamente l'importanza relativa di ogni singolo token rispetto a tutti gli altri elementi che compongono la frase o il documento. I principali vantaggi ingegneristici derivanti da questa architettura includono: la possibilità di catturare relazioni a lungo raggio nel testo e di gestire in parallelo grandi quantità di dati, superando i limiti dei modelli precedenti che processavano il testo in modo sequenziale. [35]

È importante, tuttavia, mettere in luce anche i limiti che l'NLP deve ancora superare: il problema della disambiguazione, della coerenza e del superamento delle barriere culturali e linguistiche, a cui si aggiungono i problemi dovuti alla scarsa qualità dei dataset di addestramento.

L'architettura Transformer ha reso possibile lo sviluppo dei *Large Language Models* (LLM) che hanno reso popolare l'Intelligenza artificiale Generativa (GenAI).

I LLM sono modelli di apprendimento automatico basati su architetture neurali profonde e addestrati su enormi quantità di dati testuali, da cui apprendono rappresentazioni complesse del linguaggio. Tali modelli sono emersi come sistemi di intelligenza artificiale all'avanguardia in grado di elaborare e generare testo con comunicazione coerente.[36]

L'architettura Transformer viene addestrata tramite *backpropagation* e varianti della discesa del gradiente stocastica, che aggiornano i pesi della rete minimizzando la funzione di perdita, la *cross-entropy loss*. Le *residual connections* consentono al gradiente di fluire senza degradarsi attraverso gli strati della rete, mentre i *normalization layers* stabilizzano i valori numerici interni, accelerando la convergenza e riducendo il rischio di instabilità.

Dopo il pre-training su dataset estesi, i modelli possono essere adattati a compiti specifici tramite *fine-tuning*, sfruttando dataset più piccoli e mirati.

Un Transformer si basa su un'architettura encoder-decoder[37]:

- Encoder: riceve un input testuale e genera in output una sua rappresentazione numerica vettoriale chiamata "embedding".
  Ogni strato che lo compone contiene un *self-attention layer*, per analizzare le relazioni tra le parole indipendentemente dalla loro posizione, e un *feed-forward layer*, che elabora ogni parola separatamente ed indipendentemente dalle altre. In questo livello, i dati numerici ricevuti dallo step precedente subiscono delle trasformazioni non lineari per catturare pattern complessi. Ciascuno di questi livelli è seguito da una *residual connection* e da un *normalization layer*.

  Un altro elemento essenziale è il *positional encoding*, utile a tener nota della posizione di ciascun token in modo da mantenere la struttura sequenziale del testo

- Decoder: utilizza la rappresentazione generata dall'encoder per produrre la sequenza testuale di output in modo autoregressivo. La sua struttura presenta un *self-attention layer* mascherato e un *cross-attention layer.* Il primo impedisce l'accesso ai token futuri, consentendo alla rete di valutare le relazioni tra le parole prodotte e di mantenere il contesto generato fino a quel momento. Il secondo allinea l'output con i vettori semantici dell'encoder prima del passaggio finale nel *feed-forward layer*. In questo modo, il decoder integra informazioni provenienti sia dall'input originale sia dal contesto generato fino a quel momento, permettendo al modello di produrre output coerenti e contestualmente rilevanti.

Esistono tre principali varianti di modelli basati sull'architettura Transformer: [38]

- Encoder-only: focalizzati sulla comprensione bidirezionale del testo (es. classificazione e riconoscimento di entità). Questo approccio consente al

modello di catturare in modo più efficace le sfumature semantiche e sintattiche del linguaggio naturale. A questa categoria appartengono modelli come BERT (Google) o RoBERTa (Meta), i quali nascono per analizzare il testo bidirezionalmente e vengono impiegati come potenti motori di classificazione, di ricerca semantica o nei sistemi di estrazione delle informazioni.[39], [40]

- Decoder-only: specializzati nella generazione autoregressiva di testo prevedendo il token successivo più probabile in base alla sequenza di parole già generate. Sono ideali per compiti come la scrittura creativa, i chatbot e la creazione di contenuti. Questa struttura riconduce ai modelli LLM generativi e conversazionali (come le famiglie GPT, Claude e Gemini), sistemi che, all'aumentare della scala dei parametri, hanno dimostrato un'elevata flessibilità, rendendoli lo standard.[41], [42]
- Encoder-decoder: utilizzati per compiti che richiedono sia la comprensione che la generazione di testo, come la traduzione automatica e il riassunto. Questa categoria è rappresentata da modelli quali T5 di Google o BART di Meta, architetture ibride che mappano una sequenza sorgente in uno spazio latente per poi decodificarla in una nuova sequenza target. [43], [44]

Nonostante l'elevata coerenza testuale, gli LLM presentano anche alcune limitazioni rilevanti. In primo luogo, gli LLM non sono addestrati per comprendere effettivamente il linguaggio come fa un essere umano. Infatti, come illustrato precedentemente, l'addestramento dei modelli non ha a che fare con la comprensione del significato delle parole, ma consiste nell'apprendimento delle loro associazioni statistiche, perciò soffrono del fenomeno delle "allucinazioni", ovvero la generazione di informazioni plausibili ma inesatte o prive di fondamento. Inoltre, i testi generati dipendono da dati che sono stati utilizzati durante l'addestramento. Questo costituisce un limite sia perché queste informazioni possono includere dati provenienti da fonti online non verificate e validate, sia perché tali dati sono storici e spesso non aggiornati, portando, così, a produrre

risposte obsolete o incomplete, specialmente in domini che evolvono rapidamente, come quello medico.[45] A ciò si aggiungono problematiche legate alla trasparenza e interpretabilità dei modelli, che operano come "scatole nere", rendendo difficile comprendere il processo che porta a una determinata risposta.

Infine, l'impiego degli LLM solleva questioni legate alla privacy e alla sicurezza dei dati. Sul piano della privacy, i dati sanitari costituiscono una categoria particolare di dati personali (art. 9 GDPR [46]). L'uso di LLM offerti come servizio da fornitori terzi può comportare la trasmissione di informazioni cliniche a sistemi esterni, talvolta situati fuori dall'Unione Europea, con il rischio di fuga di dati (*data leakage*) qualora gli input vengano conservati o riutilizzati per il riaddestramento. I modelli possono inoltre memorizzare porzioni dei dati di addestramento e riprodurle nell'output. Le principali misure di mitigazione sono la minimizzazione e la de-identificazione dei dati, la valutazione d'impatto sulla protezione dei dati, gli accordi con i fornitori e, ove possibile, l'esecuzione dei modelli in locale.[47] Dal punto di vista della sicurezza informatica, una delle principali preoccupazioni è la *prompt injection*, in cui un aggressore inserisce istruzioni malevole nell'input, direttamente o tramite contenuti esterni, per indurre il modello a ignorare le istruzioni originarie. [48] Nei sistemi RAG il rischio si estende alla possibile alterazione della base di conoscenza. Il quadro normativo comprende il GDPR e l'AI Act, che adotta un approccio basato sul rischio e impone, per i sistemi ad alto rischio, requisiti di trasparenza e supervisione umana.[46], [49] Questi aspetti motivano la scelta di un'architettura ancorata a una base di conoscenza verificata e che lascia la decisione finale al medico.

Queste criticità possono costituire un limite rilevante in ambito sanitario. Da un lato, gli LLM possono fornire informazioni errate o poco accurate, con possibili conseguenze sulla salute dei pazienti; dall'altro, si rischia di ledere il diritto alla privacy dei cittadini.

2.3 Retrieval-Augmented Generation (RAG)

La rigidità nella gestione di quadri clinici non previsti, presente nei sistemi esperti, e l'inaffidabilità dei modelli di linguaggio di grandi dimensioni (LLM) hanno motivato lo sviluppo di architetture ibride capaci di coniugare la trasparenza della conoscenza strutturata con la flessibilità linguistica e generativa degli LLM. In questo scenario si colloca il paradigma della *Retrieval-Augmented Generation* (RAG), un'architettura che integra un modulo di recupero dell'informazione con un modulo generativo, con l'obiettivo di basare le risposte prodotte dall'LLM su fonti verificate ed esterne al modello stesso. [50]

A differenza dei modelli generativi, il cui output dipende dalla conoscenza parametrica appresa durante l'addestramento, con i conseguenti rischi di allucinazione e obsolescenza dell'informazione, un sistema RAG interroga una base di conoscenza esterna, aggiornabile, tracciabile e pertinente. Questo è particolarmente importante per l'assistenza sanitaria, un dominio ad alta intensità di conoscenza che richiede informazioni accurate, aggiornate e specifiche per il dominio [51]. Inoltre, incorporando dati clinici aggiornati e fonti mediche affidabili come le linee guida cliniche, un LLM può offrire una consulenza più personalizzata, suggerimenti diagnostici e di trattamento più rapidi e migliorare significativamente le risposte generate.[52]

Tale approccio consente di preservare la trasparenza tipica dei sistemi esperti, poiché è possibile ricondurre ogni affermazione generata in output alla fonte da cui è stata recuperata, e l'adattabilità a contesti clinici reali dei sistemi generativi.

2.3.1 Architettura RAG

La costruzione dell'architettura RAG si articola in tre fasi sequenziali. (Figura 4)

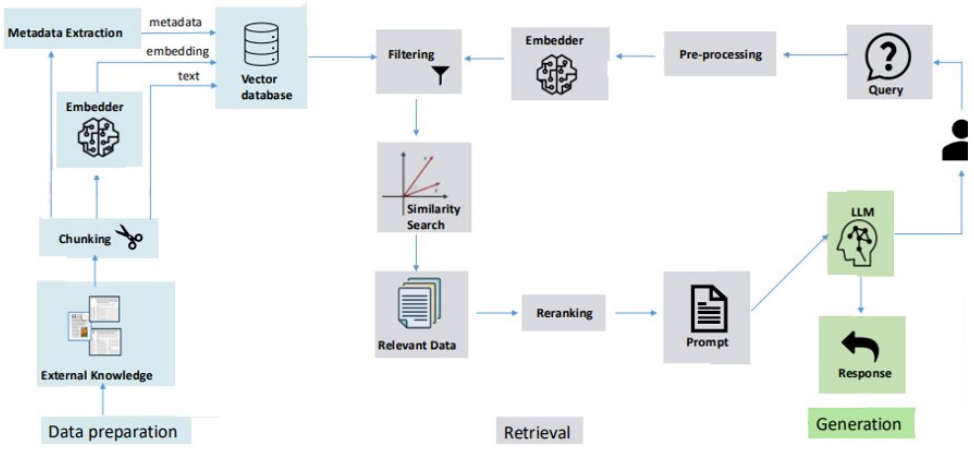


Figura 4: L'architettura Retrieval-Augmented Generation.

Il primo passo consiste nella suddivisione dei documenti in segmenti distinti, di dimensione contenuta e semanticamente coerenti (*chunk*). Nel caso del recupero denso, tali segmenti vengono trasformati in indici vettoriali tramite un modello di codifica (*encoder*) e, parallelamente, sottoposti a un processo di estrazione dei metadati. Quest'ultimo arricchisce la loro rappresentazione con informazioni strutturate aggiuntive (ad esempio la fonte, la sezione di provenienza o altri attributi rilevanti per il dominio). Testo, *embedding* e metadati vengono infine memorizzati congiuntamente all'interno di un database vettoriale, che costituisce il repository su cui si baserà la successiva fase di recupero.

In alternativa, o in combinazione con questo approccio, la conoscenza può essere strutturata sotto forma di grafo di conoscenza (*Knowledge Graph*), che rappresenta esplicitamente le entità del dominio e le relazioni semantiche che intercorrono tra esse. Questo processo, sia esso basato su indicizzazione vettoriale, su struttura a grafo o su una loro integrazione ibrida, è definito genericamente indicizzazione.

Successivamente, i segmenti o porzioni di grafo rilevanti vengono localizzati e recuperati. Nel caso del recupero denso, i segmenti vengono selezionati in base alla similarità vettoriale con la query, anch'essa sottoposta ad un processo di

embedding. Nel caso del *Knowledge Graph*, il recupero avviene tramite l'attraversamento delle relazioni semantiche codificate nella struttura del grafo oppure tramite strategie ibride che combinano i due meccanismi. I dati rilevanti così recuperati vengono infine sottoposti a un processo di disambiguazione (*reranking)*, finalizzato a riordinare i risultati secondo un criterio di pertinenza più raffinato rispetto alla sola similarità vettoriale o semantica, così da massimizzare la qualità del contesto fornito alla fase successiva.

L'ultimo passaggio prevede che il modello generi una risposta basata sul contesto derivato dagli elementi recuperati e dall'interrogazione (*query*) originale. Questi dati selezionati vengono assemblati in un prompt strutturato, che viene sottoposto all'LLM. Quest'ultimo, operando come motore generativo ancorato al contesto recuperato, elabora una risposta (*response*) fondata sulla conoscenza esterna fornita, restituendola infine all'utente.

Questi passaggi costituiscono la struttura centrale del processo RAG, rafforzando le sue capacità nel recupero delle informazioni e nella generazione di risposta consapevole del contesto. RAG sfrutta un approccio sinergico che integra il recupero delle informazioni e l'apprendimento in contesto (*In-Context Learning*: tecnica dell'LLM che consente ad esso di imparare a svolgere un nuovo compito analizzando le istruzioni fornite all'interno del testo di input[53]) per migliorare le prestazioni di un LLM. [54], [55]

2.3.2 Rappresentazione e recupero della conoscenza

Nell'architettura RAG, il recupero dell'informazione si fonda su una rappresentazione della conoscenza organizzata secondo due paradigmi complementari, spesso impiegati in combinazione. [56]

Il primo è il database vettoriale, all'interno del quale i contenuti testuali, precedentemente suddivisi in segmenti (*chunk*), vengono trasformati in rappresentazioni numeriche dense (*embedding*) tramite un modello di encoding semantico. Tra i modelli esistenti vi sono Sentence-BERT[57] e E5[58]. Il recupero

delle informazioni avviene per similarità semantica tra il vettore associato alla query e i vettori memorizzati, tipicamente calcolata tramite distanza coseno, prodotto scalare o distanza euclidea.

Dati due vettori a e b di dimensione n:

Similarità coseno:

$$\cos_sim(a,b) = \frac{a \cdot b}{||a||\,||b||} = \frac{\sum_{i=1}^{n} a_i\, b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}$$

Il risultato varia, in sistemi RAG in cui gli embedding sono tipicamente non negativi, tra 0 e 1: valori prossimi a 1 indicano un'alta similarità semantica.

Distanza coseno:

$$d_{cos} = 1 - \cos_sim(a,b)$$

Valori prossimi a 0 indicano un'alta similarità semantica.

Prodotto scalare:

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

La similarità semantica si ritrova massimizzando il risultato.

Distanza euclidea:

$$d(a,b) = ||a - b|| = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Valori prossimi a 0 indicano un'alta similarità semantica.

Questo approccio consente di individuare contenuti concettualmente affini anche in assenza di corrispondenza lessicale esatta, ma presenta il limite di non esplicitare in modo strutturato le relazioni logiche e gerarchiche tra i concetti recuperati. [59], [60]

Il secondo paradigma è rappresentato dall'utilizzo di grafi di conoscenza, una struttura dati che modella esplicitamente le entità del dominio e le relazioni semantiche che intercorrono tra esse, spesso ancorate a terminologie e ontologie standardizzate. Rispetto al database vettoriale, il grafo di conoscenza offre una rappresentazione esplicita, interpretabile e navigabile della struttura logica del dominio, consentendo il recupero di percorsi relazionali tra concetti distanti nel testo ma semanticamente connessi. [61]

L'integrazione delle due rappresentazioni consente di combinare la capacità di recupero per similarità semantica del database vettoriale con la struttura relazionale esplicita e tracciabile del grafo di conoscenza.

2.3.3 L'utilizzo dei Large Language Models come motore

Nell'architettura RAG, l'LLM assume il ruolo di motore, di gestione e generazione, sostituendo il motore d'inferenza dei sistemi esperti tradizionali. Le sue funzioni principali sono molteplici: l'interpretazione della richiesta dell'utente e la sua riformulazione in query efficaci per i moduli di recupero dell'informazione; la sintesi e l'integrazione delle informazioni recuperate dalle diverse fonti di conoscenza in una risposta coerente; l'eventuale gestione di strumenti esterni per l'esecuzione di compiti specifici non risolvibili tramite semplice generazione testuale.

A differenza del motore di inferenza di un sistema esperto, deterministico e vincolato a regole esplicite, l'LLM opera secondo un paradigma probabilistico. Questo introduce un margine di incertezza non eliminabile del tutto, ma che l'architettura RAG mira a contenere legando la generazione alla conoscenza [35], [62].

3 Rappresentazione della conoscenza

3.1 Ontologie

Nelle scienze dell'informazione, il termine ontologia viene definito come una "specificazione esplicita di una concettualizzazione" [63].

Questa definizione può essere articolata in quattro elementi fondamentali: [64]

- Formalità: un'ontologia è formale in quanto costruita secondo regole esplicite, basate su logiche formali o linguaggi standardizzati, che garantiscono precisione, coerenza e interoperabilità nella rappresentazione della conoscenza.
- Specificazione esplicita: un'ontologia fornisce una descrizione chiara e non ambigua dei concetti, delle loro proprietà e delle relazioni che li collegano.
- Condivisione: un'ontologia fornisce una conoscenza consensuale e condivisa.
- Concettualizzazione: un'ontologia rappresenta un modello astratto di un dominio, individuandone i concetti fondamentali e le interconnessioni, indipendentemente dalle occorrenze specifiche. In questo modo cattura la struttura essenziale e la semantica di un ambito di conoscenza.

La struttura di un'ontologia si basa su diversi elementi fondamentali: [64]

- Classi: categorie di concetti appartenenti al dominio di interesse, generalmente organizzate in tassonomie (concetti tra loro connessi dalla relazione is_a);
- Proprietà: descrivono le caratteristiche intrinseche della classe;
- Relazioni: rappresentano le relazioni che intercorrono tra le classi del dominio;
- Restrizioni: definiscono e delimitano il tipo di dato che la proprietà può assumere;

- Assiomi: modellizzano in maniera esplicita alcune espressioni vere in ogni caso;
- Istanze o individui: entità specifiche appartenenti alle classi. Esse ereditano tutte le proprietà della classe di cui fanno parte.

Un'ontologia è quindi una rappresentazione formale ed esplicita di un modello concettuale di dominio, che fornisce un quadro strutturato per la condivisione, l'integrazione e il ragionamento sui dati, sia all'interno di un singolo contesto sia tra sistemi eterogenei.

La struttura dell'ontologia si presta in maniera naturale ad una rappresentazione attraverso un grafo semantico, in cui i termini costituiscono i nodi e le relazioni logiche gli archi. [65] Questa struttura si integra naturalmente con i *Knowledge Graph*, poiché consente di modellare in maniera ricca e flessibile la conoscenza clinica, facilitando operazioni come la ricerca semantica, la navigazione tra concetti correlati, il supporto decisionale clinico e l'interoperabilità tra sistemi sanitari eterogenei.

3.2 Knowledge Graph

Un *Knowledge Graph* (KG), o grafo di conoscenza, costituisce una rappresentazione strutturata delle informazioni nella quale i dati vengono modellati come una rete di entità, attributi e relazioni.

Nel *Knowledge Graph* i nodi rappresentano le entità, gli archi le relazioni tra di esse e le proprietà consentono di memorizzare informazioni aggiuntive. In questo senso, il KG permette di trasformare un modello teorico della conoscenza in una struttura navigabile e interrogabile, rendendo possibile il recupero e l'analisi semantica dei dati. L'integrazione di ontologie nei KG rende quindi possibile non solo la consultazione dei dati, ma anche l'interoperabilità tra sistemi sanitari differenti.

L'unità fondamentale di un KG è la tripla (Figura 5), che esprime una singola informazione come una relazione tra tre elementi: soggetto, predicato e oggetto. Il

soggetto e l'oggetto rappresentano le entità (i nodi del grafo) di partenza e di arrivo, mentre il predicato definisce la relazione (arco) che li collega. Questo formato, standardizzato dal *Resource Description Framework* (RDF), è il fondamento per la creazione di basi di conoscenza interoperabili e semanticamente ricche. [61]

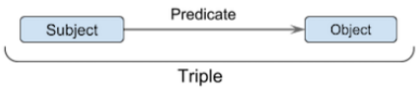


Figura 5: Tripla, unità del Knowledge Graph

Ciascuna tripla rappresenta un'informazione elementare che, combinata con altre, contribuisce alla costruzione di una rete di conoscenza: grazie a questa organizzazione, è possibile esplorare e interrogare i dati in maniera dinamica e flessibile. La natura granulare delle triple facilita l'interoperabilità tra sistemi differenti, poiché rende possibile l'integrazione di basi di conoscenza eterogenee mediante il collegamento di entità condivise attraverso relazioni formalmente definite.

Dal punto di vista implementativo vanno distinti tre livelli. L'ontologia definisce lo schema del dominio (classi, relazioni, restrizioni); il KG lo popola con le istanze o entità concrete; il database a grafo funge da infrastruttura che memorizza e interroga il grafo, in cui ogni concetto e relazione è un nodo o un arco. Rispetto ai tradizionali database tabellari, un database a grafo non si limita a organizzare i dati in forma rigida e bidimensionale, ma li struttura in modo da evidenziare le connessioni semantiche e le interdipendenze tra gli elementi. Grazie a questa struttura, diventa possibile navigare tra concetti correlati, eseguire query semantiche complesse e integrare i dati in applicazioni avanzate, come sistemi di supporto decisionale, raccomandazioni personalizzate o architetture conversazionali basate su RAG, garantendo al contempo scalabilità e prestazioni elevate su grandi quantità di dati interconnessi.

Durante il lavoro è stato utilizzato come database a grafo Neo4j, un *graph database* nativo basato sul modello *property graph*. In tale modello le entità sono

rappresentate da nodi, identificati da un'etichetta che ne identifica il tipo, e collegate da archi; sia i nodi che gli archi possono essere arricchiti da proprietà, ossia coppie chiave-valore che ne descrivono gli attributi.[66] Il grafo è interrogabile tramite il linguaggio Cypher.

3.3 Knowledge Graph per la rappresentazione di conoscenza medica

Un KG è una struttura generale: il suo valore in ambito clinico dipende dalle risorse di conoscenza con cui viene popolato. In questo lavoro la fonte principale è il testo delle linee guida per la pratica clinica, suddiviso in paragrafi. Per ancorarlo a concetti clinici standardizzati, il grafo viene integrato con risorse internazionali, ossia terminologie e sistemi di classificazione. Questi strumenti assegnano identificatori univoci ai concetti e rendono interoperabili i dati tra sistemi diversi, come la terminologia SNOMED CT e il sistema di classificazione ATC. Pur essendo entrambe rappresentabili come nodi e archi di un grafo, le due risorse hanno natura diversa. SNOMED CT è una terminologia clinica a struttura ontologica: descrive i concetti e le relazioni semantiche che li legano, ad esempio tra una malattia e la struttura anatomica coinvolta. L'ATC è invece un sistema di classificazione: organizza i principi attivi in una gerarchia di categorie, in base all'organo su cui agiscono e alle loro proprietà terapeutiche, farmacologiche e chimiche.

Nei paragrafi seguenti vengono presentate le due risorse e l'UMLS, che permette di collegarle.

3.3.1 SNOMED CT

Un'applicazione concreta del concetto di ontologia in ambito sanitario è SNOMED CT (*Systematized Nomenclature of Medicine - Clinical Terms*), la terminologia clinica più completa a livello internazionale, utilizzata per codificare, registrare e scambiare in modo standardizzato i dati sanitari elettronici. La sua adozione risponde all'esigenza di garantire una rappresentazione strutturata, standardizzata

e interoperabile dei dati clinici, rendendoli comprensibili sia ai professionisti sanitari che ai sistemi informatici. [67]

L'ontologia SNOMED CT rappresenta concetti come malattie, procedure, farmaci, segni e sintomi e le relazioni che intercorrono tra di essi. L'uso di SNOMED CT come base per i grafi di conoscenza medici rappresenta un fattore chiave nello sviluppo di applicazioni basate su intelligenza artificiale e NLP. Ciò è particolarmente rilevante nei sistemi di tipo RAG, che richiedono un accesso rapido e preciso a fonti di conoscenza strutturata per arricchire le risposte generate dagli LLM.

3.3.2 ATC

Un ulteriore sistema di classificazione rilevante in ambito clinico-farmacologico, complementare a SNOMED CT nella rappresentazione strutturata della conoscenza medica, è il sistema ATC (*Anatomical Therapeutic Chemical Classification System*), sviluppato e mantenuto dal *World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC)* dell'Organizzazione Mondiale della Sanità. A differenza di SNOMED CT, che rappresenta l'intero spettro della conoscenza clinica (malattie, procedure, segni, sintomi, farmaci e relative relazioni), l'ATC è un sistema di classificazione specificamente dedicato ai principi attivi farmacologici, organizzato secondo l'organo o l'apparato su cui il farmaco agisce e secondo le sue proprietà terapeutiche, farmacologiche e chimiche. [68]

L'ATC risulta rilevante in quanto fornisce una codifica standardizzata e interoperabile per la componente farmacologica delle linee guida cliniche. Ciò riguarda in particolare i broncodilatatori (LABA, LAMA) e i corticosteroidi inalatori (ICS) discussi nel percorso terapeutico GOLD della BPCO. Tale codifica può così integrarsi con la rappresentazione a grafo di conoscenza costruita a partire dai concetti SNOMED CT.

3.3.3 UMLS

Il collegamento tra SNOMED CT e ATC è stato definito tramite lo Unified Medical Language System (UMLS).[69] L'UMLS riunisce in un'unica risorsa molti vocabolari biomedici ed è composto da tre elementi: il Metathesaurus, la Semantic Network e lo SPECIALIST Lexicon.

Il Metathesaurus è l'elemento rilevante per l'interoperabilità tra le terminologie. Ogni termine di una fonte, con il suo codice (detto atomo), viene raggruppato con i termini delle altre fonti che hanno lo stesso significato. Il gruppo costituisce un concetto, identificato da un codice univoco, il CUI (Concept Unique Identifier). Il CUI identifica una singola idea clinica indipendentemente da come sia codificata nei diversi vocabolari: due codici di fonti diverse che condividono lo stesso CUI descrivono quindi lo stesso concetto, e il CUI funge da ponte tra i vocabolari. L'UMLS è consultabile tramite le API REST dei servizi UTS (UMLS Terminology Services)[70], che offrono la ricerca di un concetto, il recupero dei suoi atomi e una funzione di mappatura che, dato il codice di una fonte, restituisce i codici di altre fonti che condividono lo stesso CUI. Per ciascun concetto SNOMED CT corrispondente a un principio attivo si è ricavato il CUI e se ne sono cercati i codici ATC associati. Uno stesso concetto può corrispondere a più codici ATC.

3.4 Vector DB

Un database vettoriale (*vector database*) è un sistema di gestione dei dati progettato specificamente per la memorizzazione, l'indicizzazione e il recupero efficiente di dati rappresentati come vettori numerici ad alta dimensionalità, detti *embedding*. A differenza dei database relazionali tradizionali, in cui il recupero delle informazioni avviene tramite corrispondenza esatta o condizioni logiche su valori discreti, un database vettoriale organizza i dati in uno spazio vettoriale continuo, all'interno del quale la vicinanza geometrica tra due punti riflette la similarità semantica tra i contenuti che essi rappresentano.

Il presupposto teorico alla base di questo paradigma risiede nella capacità dei modelli di *embedding*, reti neurali addestrate su grandi banche dati testuali, di

proiettare unità linguistiche (parole, frasi, interi documenti) in uno spazio vettoriale denso ad alta dimensionalità, in cui contenuti semanticamente affini vengono mappati in posizioni prossime tra loro, indipendentemente dalla loro forma lessicale superficiale. (Figura 6) Questo è possibile grazie a *embedding* contestuali che sono in grado di catturare il significato delle parole in base al contesto specifico della frase in cui vengono utilizzate, consentendo il raggiungimento di rappresentazioni più sfumate e flessibili. Questa proprietà consente di superare i limiti dei tradizionali metodi di ricerca lessicale (*keyword-based*), abilitando un recupero basato sul significato piuttosto che sulla mera corrispondenza testuale.

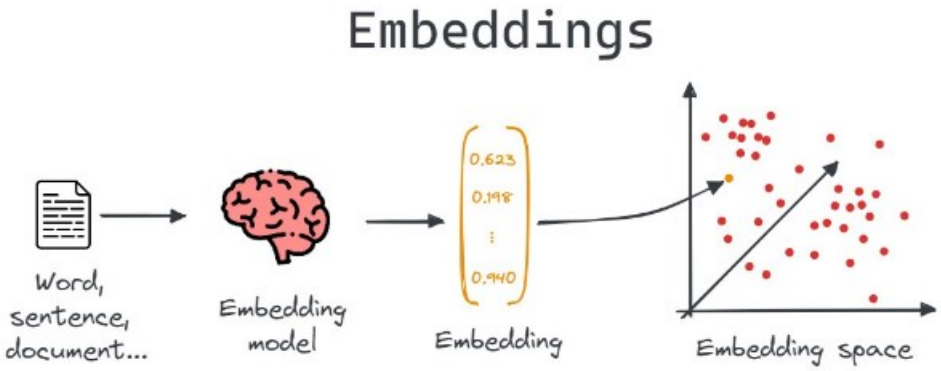


Figura 6: Rappresentazione vettoriale di dati non strutturati come quelli testuali che si dispongono in uno spazio vettoriale corrispondente [71]

Dal punto di vista strutturale, un database vettoriale organizza i propri contenuti secondo tre componenti principali: il vettore di *embedding*, che codifica numericamente il contenuto semantico dell'unità informativa; i metadati associati, che arricchiscono il vettore con informazioni strutturate aggiuntive (fonte, categoria, o altri attributi di dominio) utilizzabili per operazioni di filtraggio; e il testo originale (o un riferimento ad esso), necessario per la ricostruzione del contenuto una volta identificato come rilevante.[72]

Il recupero delle informazioni si fonda sul calcolo di una funzione di similarità (o distanza) tra il vettore associato alla query e i vettori memorizzati nello spazio

indicizzato, tipicamente tramite similarità coseno, prodotto scalare o distanza euclidea, come discusso nella sezione 2.3.2. Poiché il calcolo esatto di tale similarità rispetto a ogni vettore memorizzato (ricerca esaustiva o *brute-force*) risulta computazionalmente proibitivo su collezioni di grandi dimensioni, i database vettoriali impiegano strutture dati e algoritmi di indicizzazione per la ricerca approssimata del vicino più prossimo (*Approximate Nearest Neighbor*, ANN). Così sacrificano una minima quota di accuratezza a fronte di un significativo guadagno in termini di efficienza computazionale. [59], [73]

Nel contesto delle architetture RAG, il database vettoriale svolge dunque la funzione di infrastruttura di recupero della conoscenza non strutturata: a differenza del *Knowledge Graph*, che rappresenta esplicitamente entità e relazioni secondo una semantica formalmente definita, il database vettoriale codifica la semantica in maniera implicita e distribuita all'interno dello spazio degli *embedding*, privilegiando la flessibilità e la scalabilità del recupero a scapito dell'esplicitabilità e interpretabilità delle relazioni tra i contenuti.

4 Estrazione delle entità nel testo

Il paragrafo 3.3 ha descritto come la conoscenza clinica possa essere rappresentata in un KG: il testo delle linee guida, suddiviso in paragrafi, viene collegato alla terminologia SNOMED CT e al sistema di classificazione ATC. Per realizzare questa integrazione è necessario riconoscere nel testo i termini clinici e associarli ai corrispondenti concetti SNOMED CT. Per automatizzare questo compito si è applicato un estrattore di concetti.

La scelta di sviluppare un estrattore di concetti dedicato, anziché affidare l'intero processo all'interrogazione diretta (*prompting*) di un LLM generalista risponde a esigenze di natura tecnica e metodologica. Sebbene gli LLM abbiano dimostrato notevoli capacità di comprensione linguistica, il loro utilizzo in modalità *zero-shot* (ovvero l'interrogazione del modello tramite istruzioni testuali descrittive senza il supporto di esempi dimostrativi nel prompt[53]) per l'estrazione sistematica di concetti presenta tre limiti. In primo luogo, la generazione è probabilistica: input identici possono infatti produrre risposte differenti, parafrasi non standardizzate o "allucinazioni" terminologiche non ancorate al testo d'origine. In secondo luogo, la terminologia SNOMED CT comprende centinaia di migliaia di concetti e non può essere fornita al modello in un prompt, poiché un LLM può elaborare in una richiesta solo un numero limitato di token. Infine, una risposta generata da LLM non è riconducibile a una fonte verificabile, il che contrasta con il principio di ancorare il sistema ad una base di conoscenza. Un estrattore affronta questi limiti vincolando l'estrazione a una terminologia di riferimento. I candidati provengono da un dizionario SNOMED CT e da un indice vettoriale costruito su di esso. Ogni concetto restituito esiste quindi nella terminologia ed è riconducibile a una corrispondenza lessicale o ad un candidato recuperato, con la relativa posizione nel testo sorgente. Le componenti di estrazione e recupero sono deterministiche. L'unica variabilità residua riguarda il riordinamento con LLM (discusso al paragrafo 4.4), in cui il modello sceglie tra i candidati già recuperati. [74]

4.1 Architettura del Concept Extractor

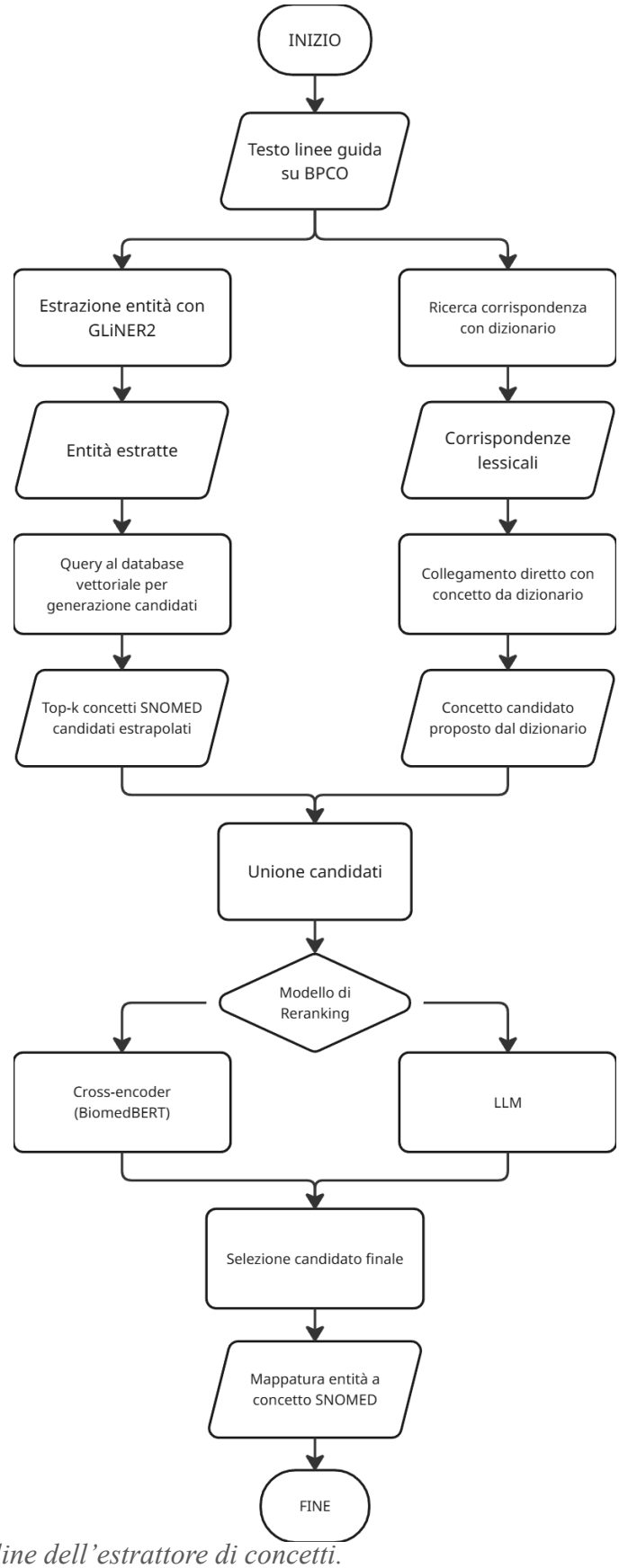


Figura 7: Pipeline dell'estrattore di concetti.

La figura 7 illustra graficamente l'architettura dell'estrattore di concetti. Il sistema adotta un approccio ibrido nella generazione dei candidati, combinando due strategie complementari di estrazione: approccio basato su dizionario (descritto nel paragrafo 4.2) e approccio neurale (descritto nel paragrafo 4.3).

I candidati prodotti dai due percorsi vengono uniti in un unico insieme, sottoposto alla fase successiva di disambiguazione. Il modulo finale di disambiguazione e riordinamento viene implementato da due modelli alternativi: *cross-encoder* ed LLM.

Entrambi gli approcci di riordinamento portano alla selezione di un candidato finale per ogni termine. Sulla base di questa estrazione automatica gli sviluppi futuri del progetto prevederanno la mappatura dei nodi del grafo di conoscenza della linea guida con i concetti SNOMED CT in essa contenuti.

4.2 Approccio basato su dizionario

L'approccio basato su dizionario costituisce il primo dei due percorsi di estrazione candidati integrati nell'architettura del sistema. Tale approccio si fonda sulla ricerca diretta, all'interno del testo clinico non strutturato, di corrispondenze lessicali con un dizionario terminologico costruito a partire dai concetti SNOMED CT del dominio di interesse. A differenza dell'approccio neurale, discusso nel paragrafo 4.3, l'individuazione di corrispondenze utilizzando il dizionario non richiede l'elaborazione degli *embedding* né l'interrogazione di un indice vettoriale. Questo approccio risulta quindi computazionalmente più leggero e immediatamente interpretabile: ogni corrispondenza individuata è infatti direttamente riconducibile al termine di dizionario che l'ha generata, senza le incertezze tipiche del recupero per similarità semantica.

L'approccio basato su dizionario, però, presenta alcune limitazioni: [76]

- incapacità di trattare flessioni e sinonimi dei termini: la ricerca individua solo le forme lessicali presenti nel dizionario, per cui varianti morfologiche e sinonimi non registrati non vengono rilevati;

- dipendenza dalla copertura del dizionario: un termine assente dal dizionario non può essere individuato. Il dizionario va inoltre aggiornato a ogni nuova versione della terminologia di riferimento;
- assenza di sensibilità al contesto: ogni corrispondenza è riconosciuta sulla sola base della forma testuale, senza considerare il significato che il termine assume nella frase. Stringhe identiche con significati diversi, come gli acronimi omografi di parole comuni (es. “LAMA”, classe farmacologica, e “lama”, organismo animale), possono quindi generare falsi positivi. Tale limite è invece ampiamente superato dall’approccio neurale (discusso al paragrafo 4.3), che opera su rappresentazioni dipendenti dal contesto;
- difficoltà con espressioni parafrasate o descrittive: un concetto formulato in modo discorsivo, anziché con il termine canonico (es. “difficoltà a respirare” in luogo di “dispnea”), non trova alcuna corrispondenza nel dizionario.

Alcune di queste limitazioni sono attenuate dalla corrispondenza approssimata (paragrafo 4.2.2), che recupera le differenze ortografiche minori (errori di battitura, piccole variazioni di forma). Essa non è, però, in grado di cogliere equivalenze semantiche tra termini lessicalmente distanti, e introduce a sua volta il rischio di falsi positivi, da contenere tramite una soglia di similarità minima. Queste limitazioni motivano l’integrazione dell’approccio basato sul dizionario con un percorso neurale (discusso nel paragrafo 4.3), capace di operare sul contesto semantico del testo anziché sulla sola forma superficiale.

4.2.1 Corrispondenza lessicale esatta

Il primo livello dell'approccio basato su dizionario consiste nell'individuazione di corrispondenze lessicali esatte tra le stringhe presenti nel testo e i termini contenuti nel dizionario terminologico. Tale operazione è stata implementata tramite l’algoritmo *FlashText* che consente di individuare in un'unica scansione lineare del documento tutte le occorrenze dei termini di dizionario, indipendentemente dal loro numero. [77]

Un aspetto metodologicamente rilevante nell'implementazione della corrispondenza esatta riguarda la gestione della sensibilità al maiuscolo/minuscolo (*case sensitivity*). Sebbene la normalizzazione del testo in minuscolo rappresenti una pratica comune nell'elaborazione testuale, tale strategia risulta problematica nel dominio clinico-farmacologico. In questo dominio, numerosi termini di rilievo sono espressi tramite acronimi brevi, la cui interpretazione dipende criticamente dalla capitalizzazione. È il caso, ad esempio, dell'acronimo LAMA (*long-acting muscarinic antagonist*), che, normalizzato in minuscolo ("lama"), collide con un termine del lessico comune privo di alcuna pertinenza clinica. Per tale ragione, la corrispondenza esatta è stata configurata per operare in modalità *case-sensitive* sui termini identificati come acronimi, preservando così la correttezza semantica dell'estrazione.

4.2.2 Corrispondenza lessicale approssimata

Il secondo livello dell'approccio basato sul dizionario è costituito dalla corrispondenza lessicale approssimata (*fuzzy matching*) introdotta per intercettare le corrispondenze che la corrispondenza esatta non è in grado di individuare a causa di variazioni lessicali minori tra il termine presente nel testo e la forma canonica registrata nel dizionario. Tra le variazioni si trovano errori tipografici, forme flesse, abbreviazioni non standard o minime difformità ortografiche. Tale funzionalità è stata implementata tramite la libreria *RapidFuzz*, che calcola un punteggio di similarità tra stringhe basato su metriche di distanza di edit (tipicamente riconducibili alla distanza di Levenshtein), esprimendo il grado di somiglianza tra due sequenze testuali come funzione del numero minimo di operazioni di inserimento, cancellazione o sostituzione di caratteri necessarie per trasformare una stringa nell'altra.[78]

La corrispondenza approssimata è stata subordinata all'introduzione di una soglia di similarità minima, al di sotto della quale una corrispondenza non viene accettata come candidato valido, al fine di contenere il rischio di falsi positivi. Nella fase di pre-processing del testo, propedeutica sia alla corrispondenza esatta sia a quella

approssimata, è stata inoltre effettuata una fase di rimozione delle *stopword,* come articoli, preposizioni, congiunzioni e pronomi (tramite la libreria NLTK), volta a eliminare dal confronto lessicale i termini funzionali privi di valore semantico specifico, riducendo così il rumore nel calcolo della similarità testuale. [79]

4.3 Approccio neurale

Accanto al percorso basato sul dizionario descritto nel paragrafo precedente, il sistema integra un secondo percorso di estrazione dei candidati, fondato su tecniche di NLP che sfruttano i modelli linguistici. Se la corrispondenza lessicale, esatta o approssimata, si dimostra efficace nell'intercettare le menzioni testualmente prossime alla forma canonica registrata nel dizionario terminologico, essa risulta limitata nel riconoscere concetti espressi con formulazioni linguistiche distanti dal termine standard: parafrasi, costrutti sintattici complessi, o espressioni implicite del concetto clinico non riconducibili ad alcuna variante lessicale pre-codificata. Colmare questo limite richiede un approccio in grado di operare non sulla forma superficiale del testo, bensì sul suo contenuto semantico.

A tal fine, l'approccio neurale si articola in tre fasi sequenziali e concettualmente distinte: (1) il riconoscimento delle entità di interesse all'interno del testo non strutturato, tramite un modello di *Named Entity Recognition* (NER, sezione 4.3.1); (2) la loro proiezione in uno spazio vettoriale semantico e il conseguente recupero dei concetti SNOMED CT candidati per similarità (sezione 4.3.2); e infine (3) il riordinamento dei candidati recuperati secondo un criterio di pertinenza più raffinato, propedeutico alla fase di disambiguazione finale (sezione 4.4).

4.3.1 Riconoscimento delle entità (NER)

Il riconoscimento di entità denominate (Named Entity Recognition, NER) è quella componente dei sistemi di NLP che si propone di individuare, all’interno di un testo, parole o frasi che fanno riferimento a una particolare entità. Per “entità denominata” si intende una parola o una frase che identifica chiaramente un elemento appartenente ad un insieme più esteso di elementi caratterizzati da

attributi simili [80]. Esempi tipici, in domini generici, sono nomi di persona, località o organizzazioni/brand; in ambito biomedico valgono come esempi nomi di malattie, farmaci o procedure. Da questi esempi emergono le due macrocategorie di entità denominate (*named entity* - NE) riconosciute dalla comunità scientifica: le NE generiche e le NE specifiche di dominio. Il NER viene spesso impiegato come fase di pre-elaborazione in varie attività, quali il recupero di informazioni, la risposta a domande o la traduzione automatica. Riconoscere le NE nelle interrogazioni aiuta a comprendere meglio l'intento dell'utente che le pone.

Si possono definire quattro filoni principali di tecniche per affrontare il compito del NER [81]:

1. Approcci basati su regole, che non necessitano di dati annotati in quanto fondati su regole predefinite;
2. Approcci di apprendimento non supervisionato, basati su algoritmi che non richiedono esempi etichettati;
3. Approcci di apprendimento supervisionato, basati su feature appositamente progettate ed etichettate e su dati annotati;
4. Approcci basati sul deep learning, che individuano automaticamente le rappresentazioni necessarie alla classificazione.

L'approccio utilizzato all'interno del progetto è quello basato sul *deep learning*. Il *deep learning* si rivela particolarmente efficace nella scoperta automatica di feature nascoste, grazie ai molteplici livelli di elaborazione che sfrutta per apprendere rappresentazioni di dati a diversi livelli di astrazione. Proprio questa composizione più complessa rispetto ai semplici modelli lineari consente l'apprendimento di caratteristiche più complesse e intricate tra i dati grezzi disponibili. In questo ambito vengono spesso impiegate architetture come le reti neurali convoluzionali (CNN) [82] e le reti neurali ricorrenti (RNN) [83] impiegate per estrapolare pattern locali e intere sequenze all'interno dei testi. Più recentemente, l'affermarsi di architetture basate su meccanismi di *attention*, i

modelli transformer, ha ulteriormente esteso le capacità dei sistemi di NER, come nel caso del modello adottato in questo lavoro. In generale, i metodi di *deep learning* per il NER seguono un flusso di lavoro articolato in tre fasi principali: rappresentazione dei dati, codifica del contesto e decodifica delle entità.

Nel progetto viene utilizzato GLiNER2, un modello zero-shot multi-task basato su un'architettura Transformer con encoder bidirezionale, progettato per unificare in un unico sistema i compiti di riconoscimento di entità, classificazione testuale ed estrazione strutturata gerarchica. GLiNER2 non richiede un insieme di etichette fisso definito in fase di addestramento, ma riceve le categorie di entità di interesse come prompt testuale al momento dell'inferenza.[75], [84] Questa caratteristica si rivela utile all'interno del progetto poiché i concetti clinici rilevanti possono essere specificati direttamente come etichette, senza la necessità di costruire un dataset annotato dedicato.

4.3.2 Rappresentazione vettoriale e recupero

Una volta individuate le menzioni testuali tramite GLiNER2, il sistema procede alla loro proiezione in uno spazio vettoriale semantico, al fine di identificare i concetti SNOMED CT candidati più affini a ciascuna entità estratta. Tale fase si fonda sui principi teorici del recupero denso già discussi nel capitolo relativo all'architettura RAG (sezione 2.3): ciascuna menzione individuata nel testo, così come ciascun concetto del vocabolario terminologico SNOMED CT, viene trasformata in una rappresentazione vettoriale densa (*embedding*) tramite un modello di *encoding* semantico, in modo tale che concetti clinicamente affini risultino prossimi tra loro nello spazio vettoriale, indipendentemente dalla loro formulazione lessicale.[57], [60]

Gli embedding dei concetti SNOMED CT del dominio di interesse vengono precalcolati e memorizzati all'interno di un indice vettoriale costruito tramite FAISS (*Facebook AI Similarity Search*), libreria impiegata per l'indicizzazione e il recupero efficiente su larga scala.[85] Al momento dell'estrazione, l'embedding di

ciascuna entità individuata da GLiNER2 viene utilizzato per interrogare tale indice, recuperando i top-k concetti SNOMED CT più simili in termini di distanza vettoriale, calcolata tramite distanza euclidea. [59]

Il recupero neurale tramite GLiNER2 ed *embedding* semantico consente di intercettare menzioni concettualmente equivalenti ma lessicalmente distanti dal termine canonico SNOMED CT. Per esempio, il termine “breathlessness”, sottoposto a recupero neurale, riporta i seguenti risultati:


```
[
      {'concept_term': ' Breathlessness', 'semantic_tag': 'finding',
'concept_id': '267036007'},
      {'concept_term': ' Has stopped breathing', 'semantic_tag': 'finding',
'concept_id': '1023001'},
      {'concept_term': ' lightheadedness', 'semantic_tag': 'finding',
'concept_id': '386705008'},
      {'concept_term': ' Cannot breathe deeply enough', 'semantic_tag':
'finding', 'concept_id': '248550001'},
      {'concept_term': 'Feeling of suffocation', 'semantic_tag': 'finding',
'concept_id': '373909009'}
]
```


L’esempio mostra come l’approccio neurale dia la possibilità di avere un ampio spettro di candidati, incrementando così la probabilità di individuare il termine più appropriato ed attinente al contesto.

I *top-k* concetti così recuperati, unitamente ai candidati individuati dal percorso basato su dizionario descritto nel paragrafo 4.2, costituiscono l'insieme di candidati sottoposto alla fase successiva di riordinamento e disambiguazione. Il recupero vettoriale restituisce un insieme di candidati ordinati per similarità semantica, ma non necessariamente pertinenti o corretti: la disambiguazione tra tali candidati rappresenta pertanto un passaggio imprescindibile per la correttezza finale della mappatura al concetto SNOMED CT appropriato.

4.4 Disambiguazione e riordinamento dei candidati

Nella fase di disambiguazione delle entità, estratte tramite i due approcci precedentemente descritti, è necessario garantire un'accurata selezione del candidato più appropriato. Per farlo, occorre tenere conto del contesto della menzione e molte strategie in letteratura propongono di adoperare una fase di riordinamento dei candidati. Questo passaggio consiste nell'assegnare uno score di rilevanza a ciascun candidato, rispetto al contesto da cui la menzione viene estrapolata. Alla fine, viene scelta l'entità con il punteggio più elevato.

Nei cross-encoder, menzione e candidato vengono elaborati assieme all'interno dello stesso modello, venendo concatenati in un'unica stringa e consentendo, dunque, di modellare in modo esplicito le interazioni tra i due testi. I risultati empirici raggiunti con questo approccio hanno portato alla definizione di quello che ad oggi è considerato lo standard nei sistemi di collegamento delle entità. Questo approccio è costoso in termini computazionali, ma al contempo molto accurato, e permette una selezione del candidato più pertinente possibile. [86]

In parallelo, i recenti sviluppi nel campo degli LLM hanno aperto nuove prospettive per la fase di riordinamento. Grazie alla loro capacità di modellizzare rappresentazioni contestuali e relazioni semantiche complesse, essi possono essere impiegati per valutare la compatibilità tra una menzione e i candidati, riformulando il problema come un compito di valutazione della similarità semantica [87]. In quest'ottica, il processo di riordinamento si sposta da una logica basata su corrispondenze lessicali ad una valutazione più profonda della comprensione del contenuto linguistico. Questo grazie all'abilità degli LLM di cogliere relazioni implicite tra diverse entità all'interno di contesti specifici. [88]

Anche in ambito biomedico tutti questi sviluppi hanno rappresentato un salto di qualità rilevante. Come già evidenziato in precedenza, questo dominio presenta una certa complessità terminologica, dovuta ad un'elevata ambiguità semantica. Servono quindi modelli in grado di cogliere le sfumature linguistiche più sottili e le relazioni concettuali più profonde. Diversi studi in letteratura confermano

questo punto: l'introduzione di moduli di riordinamento avanzati porta a miglioramenti significativi nella selezione delle entità corrette, specie nei casi più ambigui o meno frequenti.[89], [90], [91]

Coerentemente con questo duplice filone della letteratura, il sistema sviluppato in questo lavoro implementa entrambe le strategie di riordinamento appena discusse, ponendole a confronto in fase di valutazione sperimentale.

La prima si basa su un modello cross-encoder specializzato in dominio biomedico, BiomedBERT. Il modello riceve in input la coppia menzione-candidato concatenata, elaborandola congiuntamente attraverso l'intera architettura del modello. Il risultato è il punteggio di pertinenza specifico per ciascuna coppia.[92]

La seconda strategia impiega un LLM (Mistral) nel ruolo di giudice. Al modello viene sottoposto un *prompt* strutturato, con il compito di valutare e riordinare i candidati proposti rispetto alla specifica menzione intercettata nel testo. L'LLM sfrutta la propria capacità di comprensione contestuale per assegnare priorità ai candidati semanticamente più pertinenti. Il giudizio prodotto è però più variabile rispetto al punteggio esplicito del cross-encoder. Quest'ultimo è una funzione deterministica, che restituisce sempre lo stesso punteggio numerico per una data coppia menzione-candidato, perché i pesi sono fissi e non c'è campionamento. Il punteggio è continuo, quindi i candidati sono confrontabili tra loro e si può impostare una soglia. L'LLM genera la risposta campionando da una distribuzione di probabilità e può risultare sensibile alla formulazione del prompt e all'ordine in cui i candidati vengono presentati, senza fornire un punteggio calibrato e direttamente confrontabile [93]. Ne derivano una minore riproducibilità dei risultati e una minore trasparenza del criterio con cui un candidato viene preferito agli altri.

Indipendentemente dalla strategia impiegata, l'esito di questa fase è la selezione del candidato posizionato al vertice della graduatoria di pertinenza. Questo

candidato viene infine mappato come concetto SNOMED CT associato alla menzione testuale originaria, completando così il processo di estrazione.

4.5 Metriche di valutazione

La valutazione quantitativa di un sistema di estrazione automatica di concetti clinici richiede il confronto sistematico tra l'output prodotto dal sistema e un insieme di riferimento, costruito in modo indipendente, che rappresenti l'associazione corretta tra menzioni testuali e concetti SNOMED CT, definito *ground truth*.

La valutazione delle prestazioni del sistema di estrazione si fonda su un insieme di metriche standard. Data una menzione testuale, l'estrazione prodotta dal sistema può essere confrontata con la *ground truth* secondo quattro possibili esiti: un vero positivo (*true positive*, TP), quando il concetto SNOMED CT proposto dal sistema coincide con quello nella *ground truth*; un falso positivo (*false positive*, FP), quando il sistema propone un concetto non presente nella *ground truth*; un falso negativo (*false negative*, FN), quando il sistema non individua un concetto che risulta invece presente nella *ground truth*; un vero negativo (*true negative*, TN), categoria tipicamente non rilevante in compiti di estrazione in quanto indica un concetto non individuato dal sistema ed assente nella *ground truth*.

A partire da questi quattro esiti vengono definite le metriche di *precision* e *recall*.[94]

La *precision* esprime la percentuale di concetti estratti dal sistema che risultano effettivamente corretti rispetto alla ground truth, fornendo una misura di affidabilità delle estrazioni prodotte:

$$precision = \frac{TP}{TP + FP}$$

La *recall* esprime invece la proporzione di termini presenti nella ground truth che il sistema è stato in grado di individuare correttamente, fornendo una misura della completezza dell'estrazione:

$$recall = \frac{TP}{TP + FN}$$

Le due metriche sono tipicamente in tensione reciproca: un sistema configurato per massimizzare la *recall* (ad esempio, restituendo un numero elevato di candidati per ciascuna menzione) tende a penalizzare la *precision*, e viceversa.

Per sintetizzare in un unico indicatore il compromesso tra le due metriche, si ricorre comunemente alla F1-score, definita come media armonica di *precision* e *recall*:

$$F1_{score} = 2 * \frac{precision * recall}{precision + recall}$$

La scelta della media armonica, anziché di una semplice media aritmetica, risponde all'esigenza di penalizzare più severamente gli squilibri marcati tra le due componenti: la F1-score assume infatti un valore elevato solo quando sia la *precision* sia la *recall* sono contestualmente alte, risultando invece bassa qualora anche una sola delle due metriche sia scarsa, indipendentemente dal valore dell'altra.[95]

Ogni misura viene calcolata a livello del singolo frammento della linea guida: per ciascun frammento vengono infatti considerati i concetti annotati, utilizzati come *ground truth* per valutare le estrazioni prodotte dal sistema sviluppato. Successivamente, le prestazioni vengono aggregate calcolando i valori medi delle metriche, così da ottenere una valutazione complessiva dell'efficacia del sistema nell'associare correttamente i concetti ai testi.

5 Implementazione e Risultati

5.1 Pipeline

Il presente capitolo descrive l'implementazione pratica del sistema teorico presentato nei capitoli precedenti, illustrando la pipeline di lavoro complessiva seguita nello sviluppo del progetto. (Figura 8)



Figura 8: Pipeline delle fasi di implementazione del progetto di tesi

Il tema centrale del progetto di tesi è la rappresentazione della conoscenza clinica e la sua importanza nei sistemi di supporto alla decisione. Perseguire questo obiettivo richiede una formalizzazione della base di conoscenza. Questo passaggio garantisce una base di validazione rigida e affidabile, su cui ancorare e confrontare i sistemi RAG successivamente implementati.

La prima fase ha riguardato lo studio della linea guida clinica riguardante la BPCO e la conseguente formalizzazione del percorso terapeutico da essa descritto, sotto forma di *flowchart*: un passaggio propedeutico a tutte le fasi successive, poiché ha permesso di individuare in maniera esplicita e strutturata la logica decisionale contenuta nella linea guida, altrimenti espressa in linguaggio naturale discorsivo. In più, questo passaggio ha permesso di definire le variabili essenziali alla formulazione di un suggerimento clinico, identificando così le entità da estrarre come query per interrogare la base di conoscenza.

A partire da tale formalizzazione, la seconda fase ha previsto la formalizzazione della linea guida tramite il linguaggio PROforma, unitamente alla relativa fase di validazione: questo passaggio ha tradotto il percorso terapeutico, precedentemente

rappresentato come *flowchart*, in un formalismo eseguibile, coerente con l'approccio dei sistemi esperti discusso nel paragrafo 2.1.

La terza fase ha riguardato l'implementazione del *Knowledge Graph*: la suddivisione delle linee guida in *chunk*, la mappatura e la modellazione di quest'ultimi all'interno del database a grafo.

La quarta fase è consistita nell'applicazione dell'estrattore di concetti, descritto nel capitolo 4, tramite la creazione di dizionari terminologici su cui eseguire la ricerca e la costruzione della *ground truth*, necessaria alla valutazione dell'estrazione stessa. Tale passaggio è necessario a collegare, all'interno del *Knowledge Graph*, i *chunk* della linea guida ai concetti SNOMED CT in essi contenuti.

Infine, la quinta fase ha previsto il test dell'estrattore di concetti secondo le due strategie di riordinamento discusse nel capitolo 4, cross-encoder e LLM, i cui risultati sperimentali sono presentati e discussi nei paragrafi conclusivi del presente capitolo.

5.2 Flowchart

La formalizzazione del percorso terapeutico di mantenimento della BPCO, descritto nella linea guida GOLD, è stata realizzata sotto forma di *flowchart*, uno strumento di rappresentazione grafica ampiamente impiegato per esplicitare processi decisionali articolati in passaggi sequenziali e condizionali.

Prima di descrivere la macro-struttura del diagramma, è opportuno richiamare brevemente la semantica degli elementi grafici utilizzati, secondo la notazione standard adottata nella costruzione dei *flowchart*:

- il rettangolo dagli angoli arrotondati (o *stadium*) indica i punti di inizio del processo;
- il rombo rappresenta un nodo decisionale, ovvero un punto del processo in cui viene valutata una condizione il cui esito determina la diramazione del percorso in due o più rami alternativi;

- il rettangolo rappresenta un'azione o un processo da seguire;
- le frecce indicano il flusso logico e la sequenza di esecuzione tra i vari nodi.

Dal punto di vista macro-strutturale, il *flowchart* si articola in due sezioni principali, corrispondenti a due momenti distinti del percorso di cura del paziente affetto da BPCO: la sezione di avvio della terapia e la sezione di follow-up. (Figura 9)

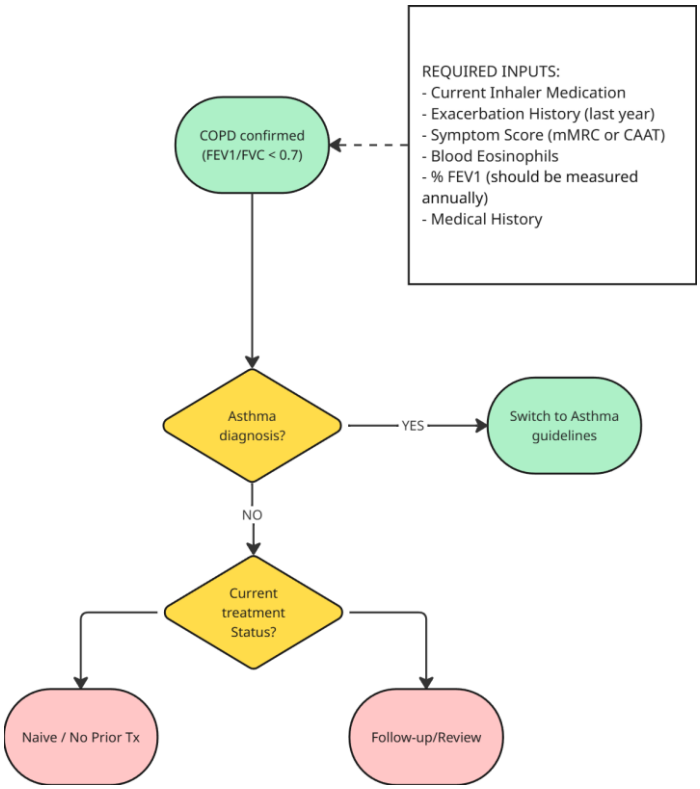


Figura 9: Flowchart delle linee guida della BPCO (prima sezione)

La prima sezione riguarda il paziente con diagnosi di BPCO confermata e FEV1/FVC < 0,7 alla prima valutazione clinica, per il quale non è ancora stata impostata una terapia farmacologica di mantenimento (paziente naive). In questa fase, il percorso decisionale conduce alla classificazione del paziente in uno dei tre gruppi di rischio previsti dalle linee guida GOLD (A, B, E), sulla base della storia di esacerbazioni nell'ultimo anno e del punteggio sintomatologico (mMRC o CAT), fino all'individuazione della terapia farmacologica iniziale più appropriata. (Figura 10)

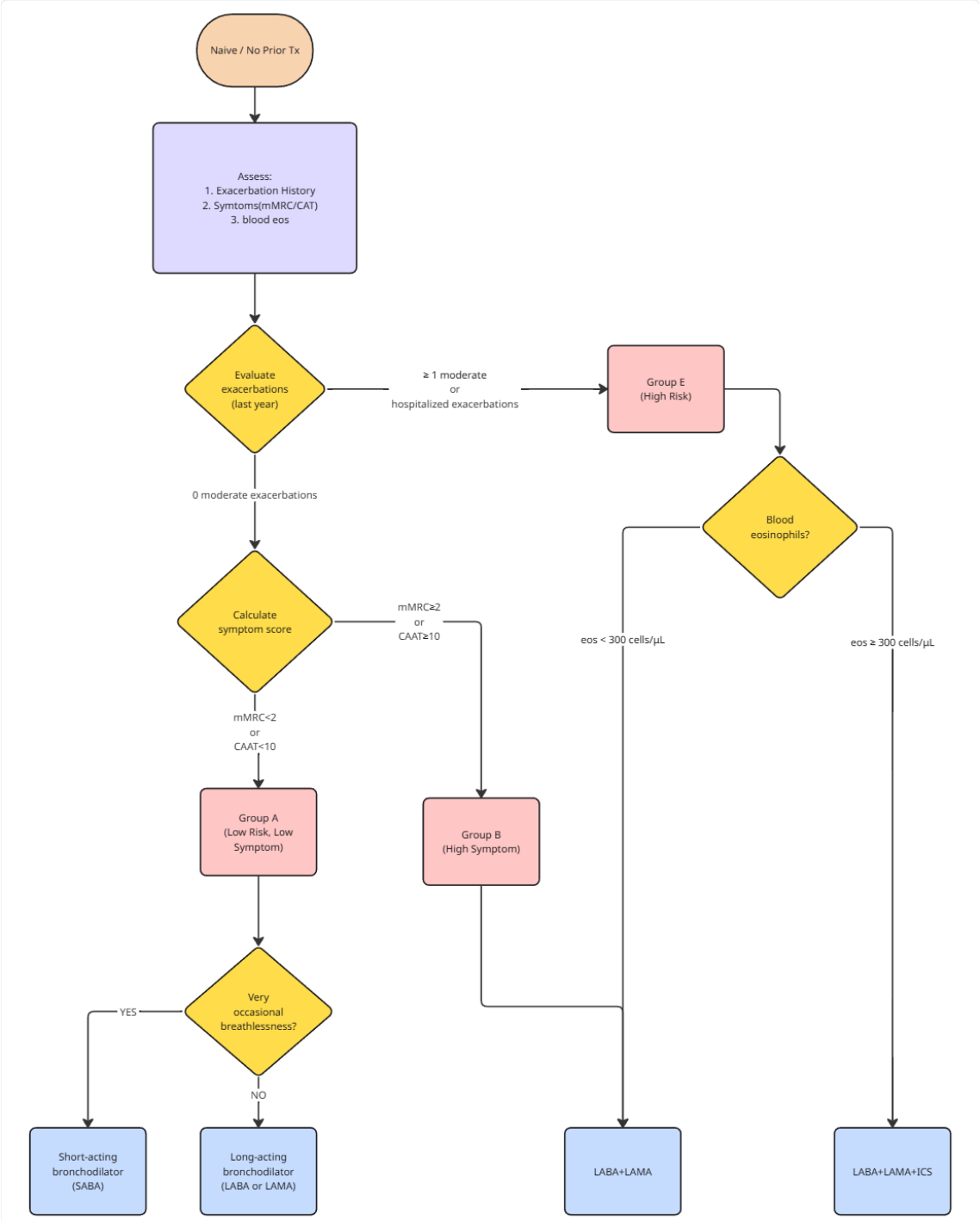


Figura 10: Flowchart delle linee guida della BPCO. Fase di inizializzazione del percorso terapeutico.

La seconda sezione, più estesa e articolata, riguarda invece il follow-up del paziente già in terapia, nel corso delle visite di controllo periodiche. (Figura 11) In questo ramo del diagramma, il percorso decisionale si sviluppa a partire dalla valutazione dell'aderenza terapeutica e della tecnica inalatoria, per poi diramarsi

in funzione della presenza di esacerbazioni persistenti (Figura 12) o di dispnea persistente (Figura 13) durante la terapia corrente. A seconda dell'esito di tali valutazioni e della conta eosinofila associata, il percorso può condurre a un'intensificazione della terapia (*step-up*), a una sua riduzione (*step-down*, ad esempio la sospensione degli ICS in assenza di beneficio clinico evidente), oppure alla conferma della terapia in corso, con eventuale combinazione con trattamenti aggiuntivi in casi clinici specifici (ad esempio azitromicina, roflumilast, o farmaci biologici come mepolizumab e dupilumab).

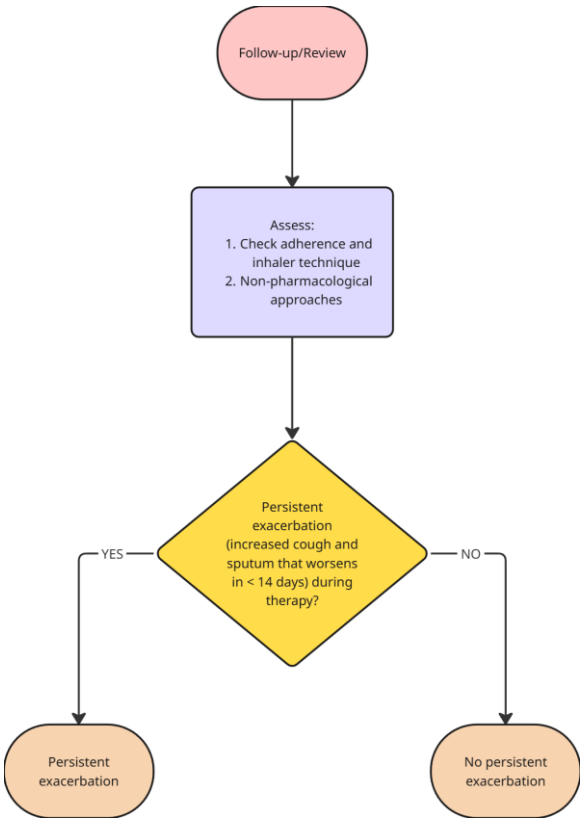


Figura 11: Flowchart delle linee guida della BPCO. Fase di revisione del percorso terapeutico.

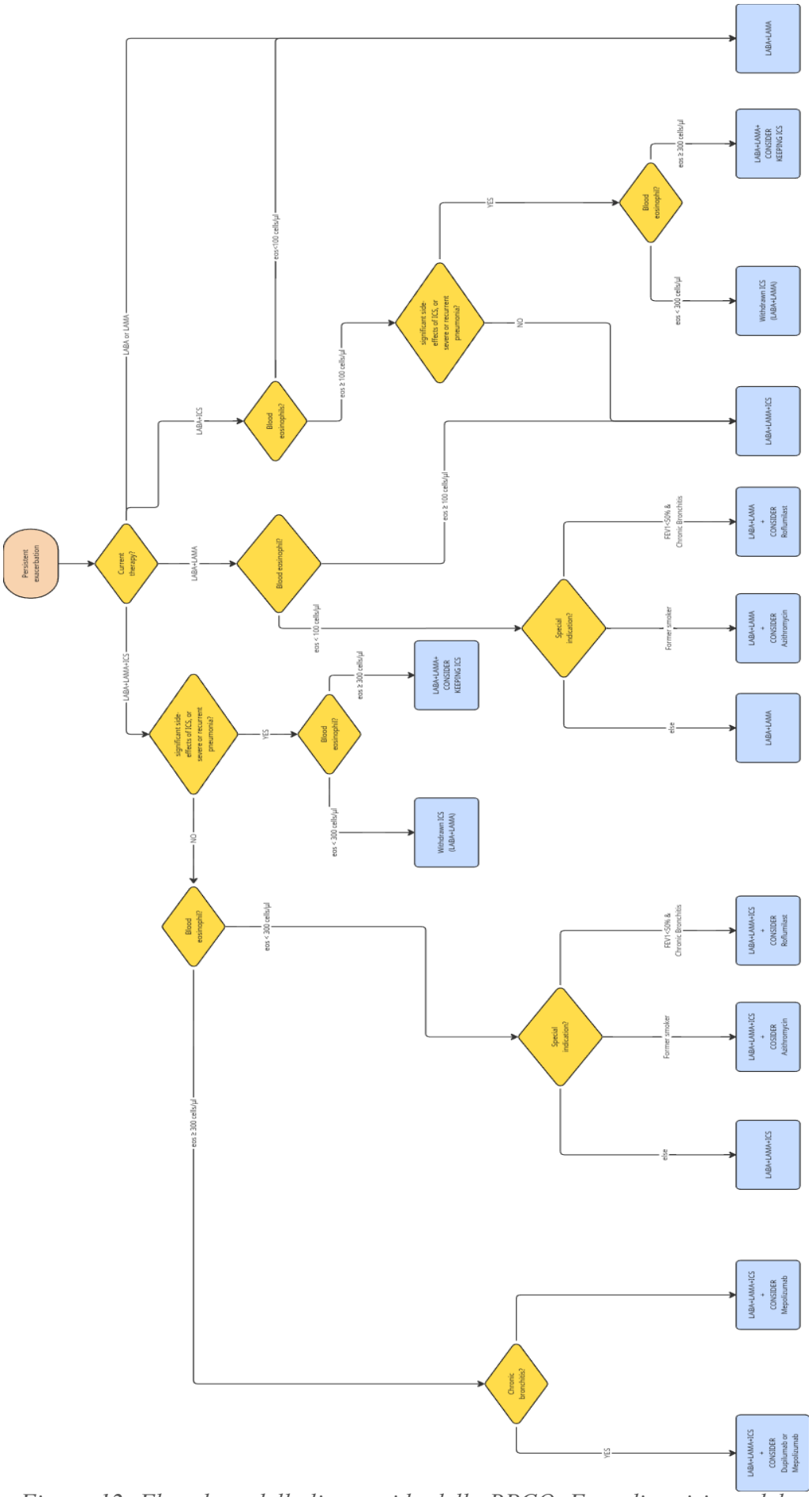


Figura 12: Flowchart delle linee guida della BPCO. Fase di revisione del percorso terapeutico. Ramo da percorrere in presenza di esacerbazioni persistenti.

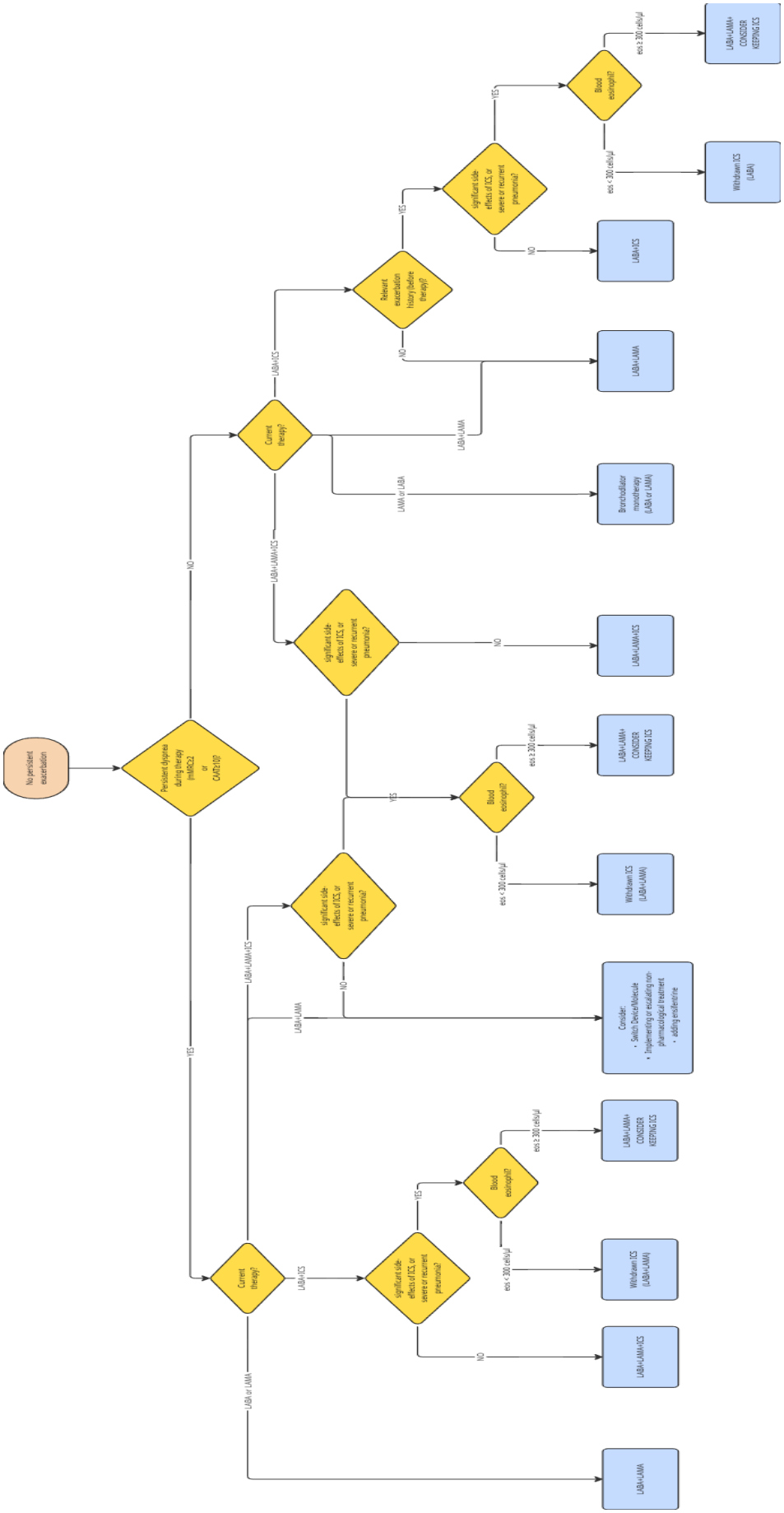


Figura 13: Flowchart delle linee guida della BPCO. Fase di revisione del percorso terapeutico. Ramo da percorrere in assenza di esacerbazioni persistenti.

Un elemento trasversale a entrambe le sezioni del *flowchart* è la logica di progressione terapeutica, articolata su tre livelli crescenti di intensità farmacologica:

- monoterapia con un singolo broncodilatatore a lunga durata d'azione (LABA o LAMA), riservata tipicamente al paziente a basso rischio e bassa sintomatologia (gruppo A) o come primo tentativo terapeutico in assenza di indicazioni più specifiche;
- duplice terapia broncodilatatoria (LABA+LAMA), indicata nel paziente con sintomatologia più marcata (gruppo B) o come primo step di escalation in caso di controllo insufficiente con la monoterapia;
- tripla terapia (LABA+LAMA+ICS), riservata al paziente ad alto rischio di esacerbazioni (gruppo E) con conta eosinofila elevata (≥300 cellule/µl), o introdotta in fase di follow-up qualora le esacerbazioni persistano nonostante la duplice terapia e il valore di eosinofili lo giustifichi (tipicamente ≥100 cellule/µl).

Tale logica di progressione, reversibile in entrambe le direzioni a seconda della risposta clinica del paziente, riflette la natura dinamica e personalizzata del percorso terapeutico previsto dalla linea guida GOLD.

La creazione del *flowchart* ha permesso di analizzare l'intero percorso terapeutico e di individuare, in collaborazione con esperti clinici, le variabili essenziali per ottenere una panoramica corretta e completa della condizione clinica del paziente. Definire con precisione tali entità cliniche è infatti condizione necessaria per poter percorrere l'intero flowchart fino ad un suggerimento terapeutico. Senza di esse, infatti, le informazioni disponibili sul paziente risulterebbero parziali, rendendo impossibile una risposta esaustiva al quesito medico.

La Tabella 2 riporta le variabili individuate, ciascuna con il proprio tipo di dato, l'unità di misura dove applicabile e l'impatto specifico sulla decisione terapeutica.

Variabili per il trattamento della BPCO (GOLD 2026)

Variabile	Tipo	Unità	Impatto sulla decisione
mMRC Dyspnea Score	Intero	0–4 scala	Punteggio ≥ 2 indica un alto grado di sintomi (Gruppo B)
CAT Score	Intero	0–40 scala	Punteggio ≥ 10 indica un alto grado di sintomi (Gruppo B)
Conta eosinofili nel sangue	Decimale	cells/µL	<100 (Evitare ICS) vs ≥300 (ICS altamente raccomandato)
FEV1 (Volume Espiratorio Forzato nel primo secondo)	Decimale	% (/ L)	
FVC (Capacità Vitale Forzata)	Decimale	% (/ L)	
Dispnea persistente (durante la terapia)	Booleano	Sì / No	
Mancanza di respiro solo occasionale	Booleano	Sì / No	
Indicazioni speciali	Lista[Stringa]	Ex fumatore / bronchite cronica / altro	
Esacerbazioni	Intero	Conta (ultimi 12 mesi)	≥ 1 definisce un alto rischio (Gruppo E)

Asma concomitante	Booleano	Sì / No	Passaggio a linea guida dell'asma
Polmonite	Booleano	Sì / No	Controindicazione per l'uso di ICS
Significativi effetti collaterali dell'ICS	Booleano	Sì / No	
Terapia corrente (dose cumulativa al giorno)	Lista[str; decimale]	Nessuna terapia precedente / [nome farmaco; mg]	
Comorbidità	Lista[str]	Comorbidità	
Eventi di esacerbazioni rilevanti (prima dell'inizio della terapia)	Booleano	Sì / No	

Tabella 2: Variabili per il trattamento della BPCO secondo l'analisi di GOLD 2026

Poiché una delle fasi del progetto prevede l'estrazione di queste variabili cliniche a partire da referti testuali, la loro definizione puntuale costituisce un passaggio essenziale, propedeutico sia alla costruzione degli strumenti di estrazione sia alla loro successiva validazione.

5.3 Modulo strutturato compilabile

L'analisi dei referti clinici forniti dagli ospedali partner ha evidenziato l'assenza di variabili necessarie al raggiungimento del suggerimento terapeutico. Si è scelto, perciò, di istruire i medici degli ospedali coinvolti nel progetto, comunicando loro

quali variabili fossero imprescindibili affinché il sistema di intelligenza artificiale potesse produrre output esaustivi. Per questo motivo è stato costruito un modulo strutturato, compilabile direttamente dai medici, le cui annotazioni costituiscono il *gold standard* rispetto al quale vengono successivamente valutati gli approcci automatici di estrazione su referti medici in linguaggio naturale. Il modulo è organizzato in quattro sezioni, illustrate di seguito tramite un esempio.

Sezione A: Criteri clinici (Figura 14)

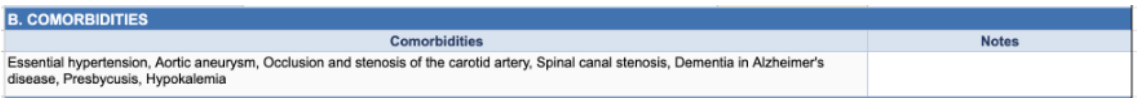
Per ciascuna variabile predefinita è necessario selezionare l'opzione appropriata da un menu a tendina, oppure inserire il dato come testo libero, specificando l'unità di misura dove necessario. Ogni variabile è corredata da una definizione e da un'indicazione clinica a supporto dell'interpretazione. Qualora l'informazione richiesta non sia riportata nel referto, la cella va lasciata vuota, accompagnata da una nota che segnali esplicitamente l'assenza.

A. CLINICAL CRITERIA | Select the answer from the dropdown if available, otherwise report the answer as free-text			
Variable	**Definition / Clinical Hint**	**Answer**	**Notes**
mMRC Dyspnea Score	Modified Medical Research Council (mMRC) Dyspnea Score (0-4) that grades the severity of breathlessness based on its impact on daily activities.	3	
CAT Score	COPD Assessment Test (CAT) Score (0-40) that quantify the overall impact of COPD on the patient's health status and daily life.	28	
Blood Eosinophil Count	Blood eosinophil count with unit of measurement.	50 /µl	
FEV1	Forced expiratory volume in 1 second with unit of measurement.	26%	
FVC	Forced vital capacity with unit of measurement.		not reported
Dyspnea	Is persistent dyspnea the leading and only trait?	TRUE	
Occasional breathlessness	Occasional breathlessness.	FALSE	
Special indication	Special indication reported or observed. Select one from the predefined list: Smoker / Chronic bronchitis / Bronchiectasis / Other.	None	
Exacerbation	True if the patient experienced an exacerbation during the observation period, defined as a worsening of cough and/or sputum production lasting less than 14 days (for patients already on therapy) or occurring within the last 12 months (for therapy-naïve patients).	TRUE	
Pneumonia history	Pneumonia history. Select one from the predefined list: No / Recurrent / Severe.	No	
ICS significant side effects	Significant side effects of inhaled corticosteroids (ICS) reported or observed. Select all that apply from the predefined list. Set to null if no significant ICS side effects are reported.	None	

Figura 14: modulo strutturato compilabile, sezione dedicata ai criteri clinici.

Sezione B: Comorbidità (Figura 15)

Si richiede di riportare tutte le condizioni di comorbidità esplicitamente menzionate nel referto, separate da una virgola.

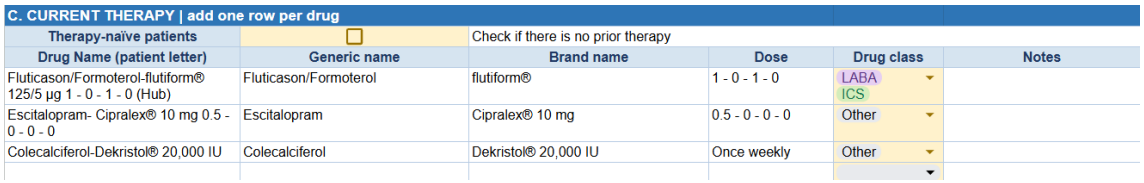


B. COMORBIDITIES	
Comorbidities	Notes
Essential hypertension, Aortic aneurysm, Occlusion and stenosis of the carotid artery, Spinal canal stenosis, Dementia in Alzheimer's disease, Presbycusis, Hypokalemia	

Figura 15: modulo strutturato compilabile, sezione dedicata alle comorbidità.

Sezione C: Terapia corrente (Figura 16)

Vanno elencati tutti i farmaci assunti dal paziente al momento della visita, con una riga dedicata a ciascun farmaco. Se il paziente non sta seguendo alcuna terapia, è possibile selezionare la voce "*therapy-naive patient*" e lasciare i restanti campi vuoti.

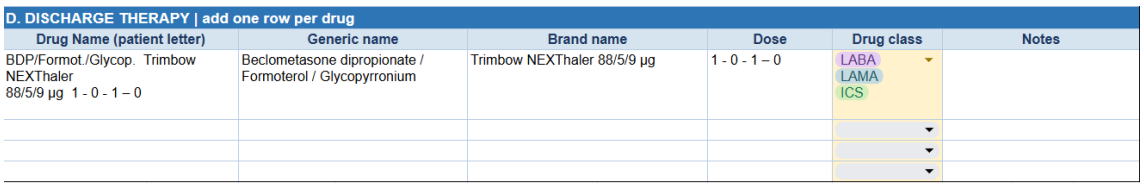


C. CURRENT THERAPY | add one row per drug					
Therapy-naïve patients	☐	Check if there is no prior therapy			
Drug Name (patient letter)	Generic name	Brand name	Dose	Drug class	Notes
Fluticason/Formoterol-flutiform® 125/5 µg 1 - 0 - 1 - 0 (Hub)	Fluticason/Formoterol	flutiform®	1 - 0 - 1 - 0	LABA ICS	
Escitalopram- Cipralex® 10 mg 0.5 - 0 - 0 - 0	Escitalopram	Cipralex® 10 mg	0.5 - 0 - 0 - 0	Other	
Colecalciferol-Dekristol® 20,000 IU	Colecalciferol	Dekristol® 20,000 IU	Once weekly	Other	

Figura 16: modulo strutturato compilabile, sezione dedicata alla terapia corrente.

Sezione D: Terapia di dimissione (Figura 17)

Vanno elencati tutti i farmaci prescritti, da continuare o iniziare a seguito della visita, con una riga dedicata a ciascun farmaco.



D. DISCHARGE THERAPY | add one row per drug					
Drug Name (patient letter)	Generic name	Brand name	Dose	Drug class	Notes
BDP/Formot./Glycop. Trimbow NEXThaler 88/5/9 µg 1 - 0 - 1 – 0	Beclometasone dipropionate / Formoterol / Glycopyrronium	Trimbow NEXThaler 88/5/9 µg	1 - 0 - 1 – 0	LABA LAMA ICS	

Figura 17: modulo strutturato compilabile, sezione dedicata alla terapia di dimissione.

In entrambe le ultime due sezioni si raccomanda di indicare il nome completo del farmaco alla voce "*Drug Name*", i principi attivi alla voce "*Generic name*" e il nome commerciale alla voce "*Brand name*". Vanno, inoltre, inserite le dosi e le classi di appartenenza dei farmaci.

5.4 Formalizzazione della linea guida tramite linguaggio PROforma

A partire dalla formalizzazione del percorso terapeutico in forma di *flowchart*, la seconda fase del progetto ha previsto la formalizzazione della linea guida tramite il linguaggio PROforma utilizzando il software Deontics Composer, unitamente alla relativa fase di validazione. Questo passaggio ha tradotto il percorso terapeutico in un formalismo eseguibile, coerente con l'approccio dei sistemi esperti mostrato nel paragrafo 2.1 e con l'ontologia dei task descritta nel paragrafo 2.1.1.

Come già discusso, PROforma è un sistema esperto in grado di raccogliere dati e formalizzare suggerimenti clinici sulla base di regole fisse e rigide, fondate su una logica deontica che ne governa il comportamento decisionale. Questo garantisce l'affidabilità del sistema. PROforma consente inoltre di definire in dettaglio ogni elemento del processo, di inserire collegamenti diretti ai documenti di riferimento, in questo caso la linea guida GOLD sulla BPCO, e di giustificare ogni scelta e suggerimento articolandone pro e contro. È infine importante sottolineare che nessun sistema di supporto decisionale clinico può sostituirsi al medico nella decisione finale: ne consegue che ogni task *decision* consente al medico di selezionare liberamente qualunque opzione, anche qualora non sia quella raccomandata dal sistema.

La formalizzazione ha il compito di raccogliere i dati sintomatologici del paziente e, sulla base di questi, suggerire il percorso terapeutico più appropriato, rimanendo fedele alle indicazioni della linea guida sulla BPCO. Il sistema è pensato per pazienti con diagnosi di BPCO già formulata, poiché la diagnosi esula dagli obiettivi del progetto: si assume, dunque, che il medico faccia ricorso al sistema solo in presenza di un quadro clinico già compatibile con tale diagnosi.

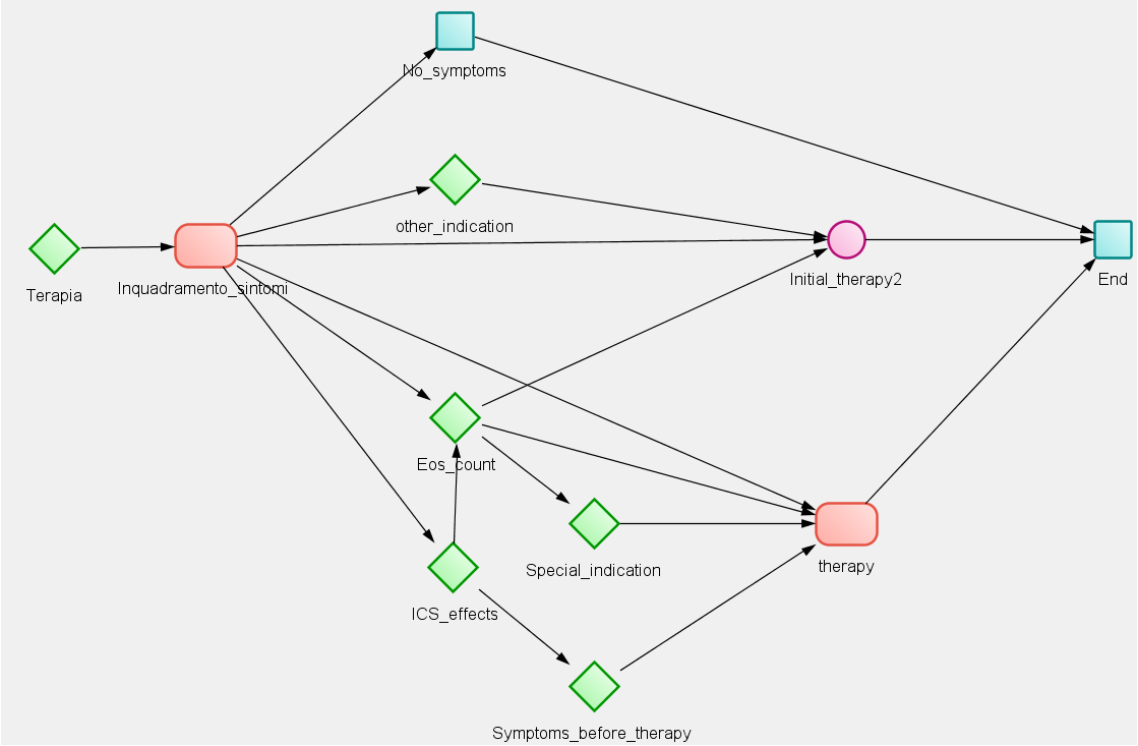


Figura 18: struttura della modellizzazione con il Composer.

Il primo passaggio, rappresentato dalla *enquiry* "Terapia" (Figura 18), prevede la definizione della situazione terapeutica in atto: il paziente potrebbe essere stato appena diagnosticato, nel qual caso il percorso si dirige verso la definizione di una terapia iniziale (*decision* "initial_therapy"). In alternativa, il paziente potrebbe già essere sotto trattamento farmacologico. In tal caso, è necessario stabilire se si tratti di mono, duplice o tripla terapia. Il percorso prevede quindi un aumento dei quesiti sintomatici e si indirizza verso la definizione di una nuova terapia, che può confermare, incrementare o ridurre quella già in corso (*plan* "therapy").

Un passaggio obbligato in qualsiasi percorso è l'esecuzione del *plan* "Inquadramento_sintomi" (Figura 19), all'interno del quale viene definita la situazione sintomatica generale del paziente. In particolare, si chiede se vi siano stati episodi di esacerbazione, moderata o acuta (risposta booleana), negli ultimi 12 mesi nel caso di un paziente *naive*, oppure durante la terapia corrente, nel caso di un paziente già in cura. Successivamente, la *decision* "Dyspnea_tools" permette di scegliere il metodo con cui quantificare il livello dei sintomi di dispnea. Le

opzioni raccomandate sono i due questionari mMRC (punteggio intero, scala 0–4) e CAT (punteggio intero, scala 0–40), considerati interscambiabili in termini di efficacia. È invece sempre sconsigliato l'inserimento di una valutazione clinica puramente soggettiva, basata sulla sola impressione del medico. Questa opzione è prevista soltanto nel caso in cui gli strumenti standardizzati non siano disponibili e rappresentino l'unico modo per raccogliere l'informazione.

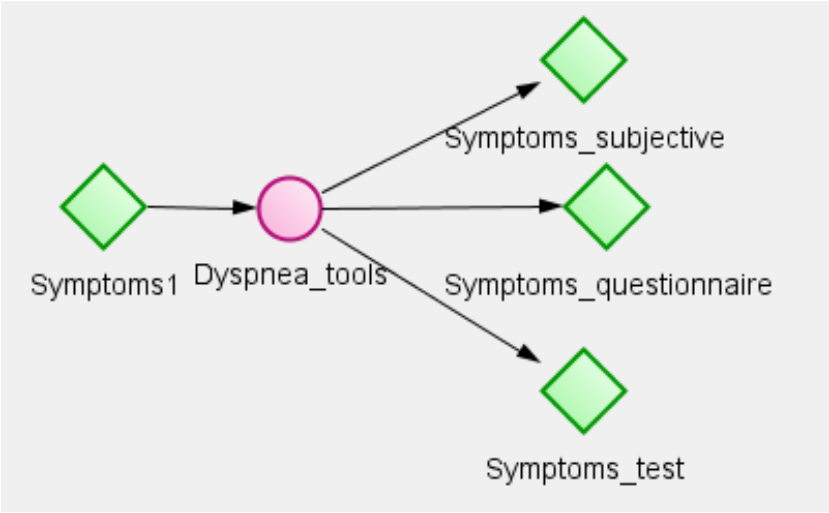


Figura 19: struttura del Plan inquadramento_sintomi

Il sistema si conclude nel caso in cui il paziente non sia sotto cura e non presenti sintomi, tramite il task *action* "No_symptom".

Una volta definiti i sintomi più direttamente legati alla patologia in esame, vengono eseguiti, se necessario e secondo quanto stabilito dalla linea guida, ulteriori accertamenti, ad esempio il conteggio degli eosinofili nel sangue o, in caso di triplice terapia, la valutazione degli effetti collaterali dell'ICS. Di seguito si definiscono i diversi task *enquiry* impiegati e le informazioni da essi richieste:

1. “other_indication”: chiede se la mancanza di respiro sia solo occasionale (risposta booleana)
2. “Eos_count”: chiede di inserire il tasso di eosinofili nel sangue, in cellule/µL (risposta numerica reale)

3. “ICS_effect”: chiede di segnalare eventuali effetti collaterali dell’ICS, tra i seguenti:
    - Riduzione della densità ossea
    - Cataratta
    - Glaucoma
    - Diabete
    - Disfonia
    - Infezione micobatterica
    - Orofaringite da Candida
    - Polmonite
    - Abrasioni cutanee
    - Tubercolosi
4. “Special_indication”: chiede di rispondere a tre diverse condizioni:
    - Bronchite cronica (risposta booleana)
    - FEV1 in % (risposta numerica reale)
    - Ex fumatore (risposta booleana)
5. “Symptoms_before_therapy”: chiede se il paziente in terapia abbia manifestato esacerbazioni rilevanti prima della diagnosi di BPCO e, quindi, prima dell’inizio di qualsiasi terapia correlata.

I nodi cardine del sistema risiedono nei task *decision*, in cui viene definita la terapia da prescrivere e il sistema giunge a compimento del proprio obiettivo di supporto alla decisione clinica. Per il paziente *naive* (mai trattato in precedenza), questo passaggio è affidato a un unico task *decision*, le cui opzioni candidate sono:

- broncodilatatori a breve durata d'azione (SABA);
- monoterapia con broncodilatatore a lunga durata d'azione: beta2-agonista (LABA) o antimuscarinico (LAMA);
- duplice terapia, combinando i due broncodilatatori a lunga durata d'azione (LABA+LAMA);

- triplice terapia, in cui ai due broncodilatatori si aggiunge un corticosteroide inalatorio (ICS).

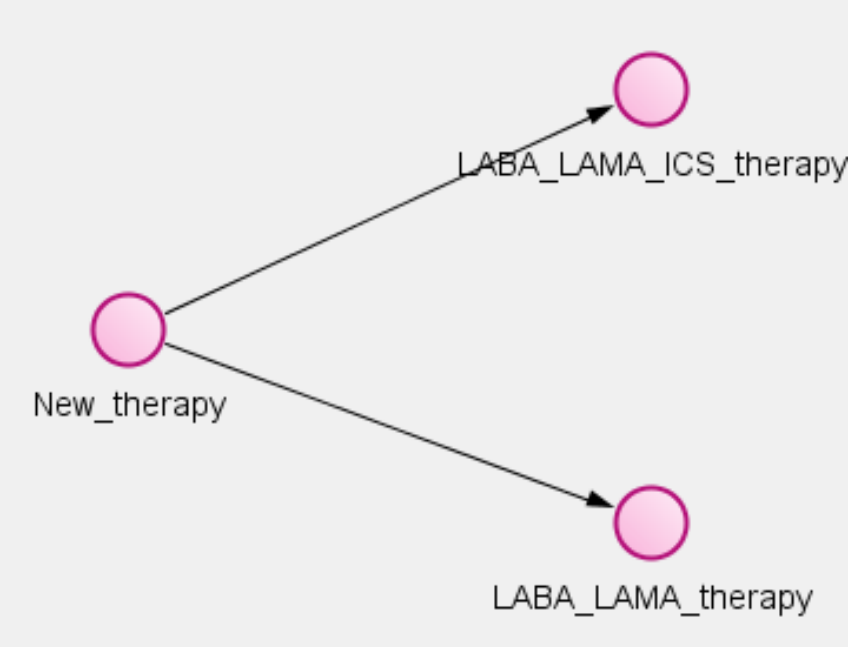


Figura 20: struttura del Plan therapy

La selezione della terapia, nel caso in cui il paziente sia sottoposto ad una visita di follow-up o di revisione, è contenuta nel *plan* "therapy" e si articola in tre task *decision* concatenati (Figura 20). Il percorso inizia con il task *decision* "New_therapy", che propone una tra le terapie standard: mono, duplice, triplice, oppure la combinazione LABA+ICS (generalmente sconsigliata dalla linea guida, ma mantenibile qualora si riveli clinicamente efficace). Qualora venga selezionata la duplice terapia e i sintomi del paziente lo richiedano, il percorso prosegue nel task "LABA_LAMA_therapy". In questo caso, alla combinazione LABA+LAMA è possibile associare Roflumilast, Azitromicina, entrambi i farmaci oppure la prosecuzione dell'ICS, se già in uso. Qualora venga invece selezionata la triplice terapia, e i sintomi lo consentano, il percorso prosegue nel task "LABA_LAMA_ICS_therapy", in cui alla terapia selezionata è possibile associare Roflumilast, Azitromicina, entrambi, oppure una terapia biologica quale Mepolizumab o Dupilumab.

Una volta completata la prescrizione, il sistema giunge al nodo terminale, il task *action* "End". In questo nodo viene segnalato il completamento dell'esecuzione della linea guida per il caso in esame e viene raccomandata la programmazione di una visita di follow-up, suggerendo le tempistiche più adeguate in base alla situazione clinica specifica del paziente.

La validazione di un sistema di supporto alla decisione clinica non può prescindere dal confronto con casi reali, tratti dalla pratica clinica effettiva. I casi clinici reali, infatti, riflettono la variabilità intrinseca della presentazione clinica dei pazienti (dati mancanti, informazioni riportate in forma non standardizzata, sovrapposizione di più condizioni cliniche) che un sistema formalizzato deve essere in grado di gestire correttamente per poter essere considerato affidabile in un contesto operativo reale. Validare il sistema esclusivamente su casi teorici, costruiti per rappresentare in modo idealizzato ogni ramo del percorso decisionale, rischierebbe di non intercettare criticità che emergono solo nell'interazione con dati clinici autentici. Tra queste criticità figurano l'ambiguità nella documentazione, la terminologia non uniforme tra referti diversi e le combinazioni di variabili non esplicitamente previste in fase di progettazione. Per questo motivo, il sistema è stato validato grazie a casi clinici forniti dagli ospedali partner del progetto; di seguito viene presentato un esempio applicativo.

Dato un referto clinico, vengono innanzitutto estratte le variabili di interesse, riportate in Tabella 3.

Variable	Ground truth
mMRC Dyspnea Score (Modified Medical Research Council Dyspnea Score [0-4])	IV
CAT Score (COPD Assessment Test Score [0-40])	none
Blood Eosinophil Count (cells/µL)	48
FEV1 (% forced expiratory volume in 1 second)	none
FVC (% forced vital capacity)	none
Persistent Dyspnea (during therapy)	yes
Occasional breathlessness	no
Special indication (Former Smoker, Chronic bronchitis, Other)	none
Exacerbation (increased cough and sputum that worsens in <14 days during therapy or in <12 months if naïve patient)	>=2
Pneumonia	no
Significant side-effects of Inhaled corticosteroids (ICS)	none
Admission/Current therapy	BDP/Formot./Glycop. Trimbow NEXThaler 88/5/9 µg 1 - 0 - 1 - 0
Discharge therapy	LABA+LAMA+ICS

Tabella 3: Variabili di interesse per la BPCO estratte da un referto clinico

Per verificare il funzionamento della linea guida formalizzata in PROforma, è stato utilizzato il Tester del sistema Deontics, che consente di eseguire il modello decisionale e osservarne il comportamento sulla base dei dati clinici inseriti. Durante l'esecuzione, il sistema presenta una sequenza di schermate interattive attraverso le quali l'utente specifica le caratteristiche cliniche del paziente, selezionando le opzioni disponibili o inserendo i valori richiesti. Le informazioni raccolte vengono utilizzate per valutare le condizioni definite nel modello e determinare le raccomandazioni terapeutiche coerenti con il percorso decisionale formalizzato.

Nel caso in esame, l'esecuzione riguarda il percorso terapeutico di mantenimento della BPCO. A partire dalla terapia farmacologica corrente, vengono acquisiti progressivamente i dati relativi alla presenza di esacerbazioni, alla gravità della dispnea, agli eventuali effetti collaterali dei corticosteroidi inalatori (ICS), alla conta degli eosinofili nel sangue e alla presenza di indicazioni cliniche specifiche. Le figure seguenti documentano le schermate del Tester corrispondenti ai diversi input richiesti e alla raccomandazione terapeutica risultante.

1. Selezionare la terapia corrente: Triplice terapia (Figura 21)

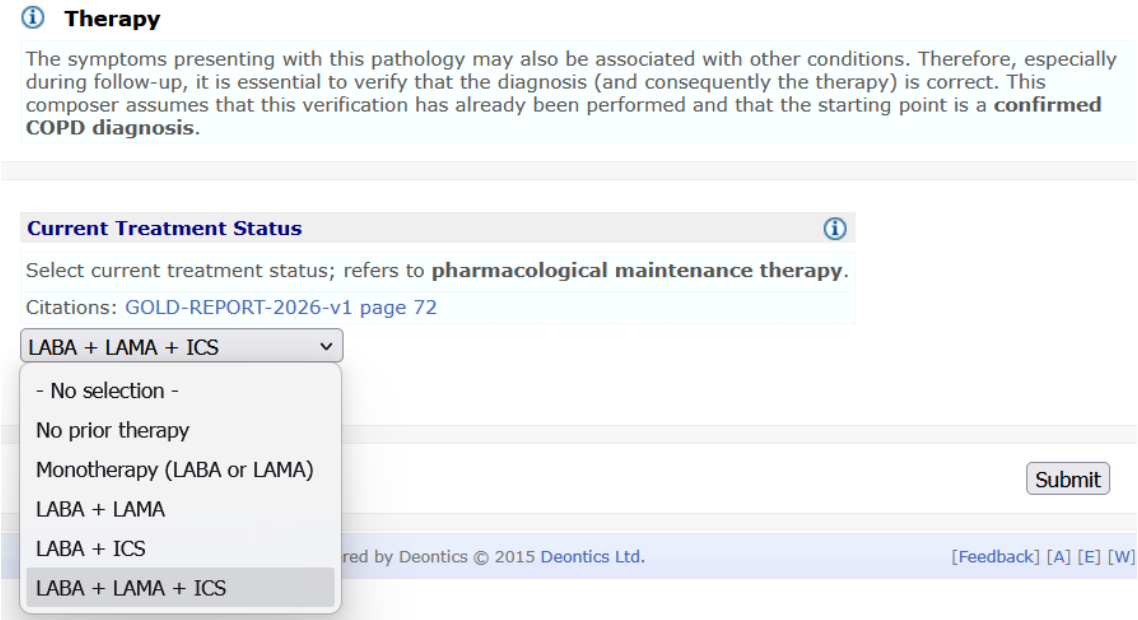


Figura 21: Tester – Selezione della terapia corrente

2. Definire la presenza di esacerbazioni: Sì (Figura 22)



Figura 22: Tester – Presenza di esacerbazioni.

3. Scegliere lo strumento di valutazione dei sintomi di dispnea: mMRC (Figura 23)
   Riportare il valore: 4 (Figura 24)

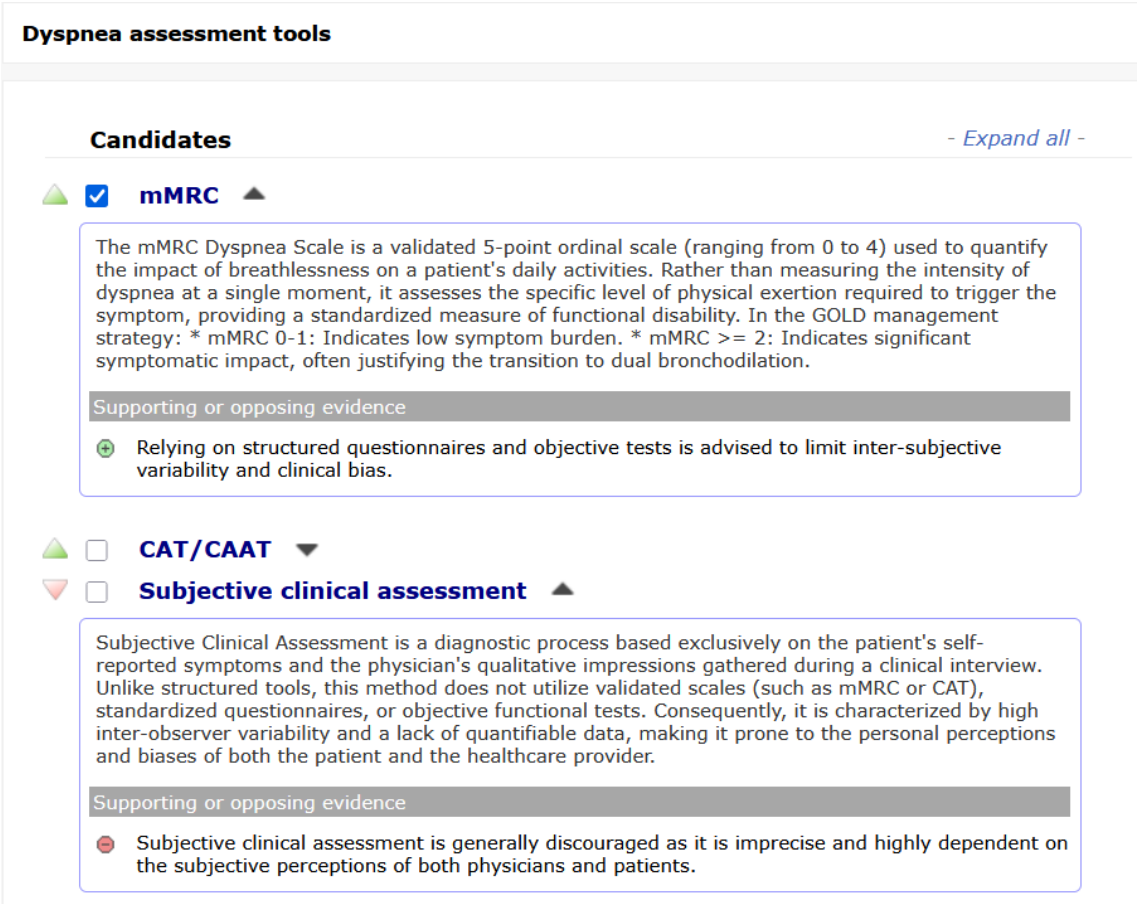


Figura 23: Tester – Scelta dello strumento di valutazione

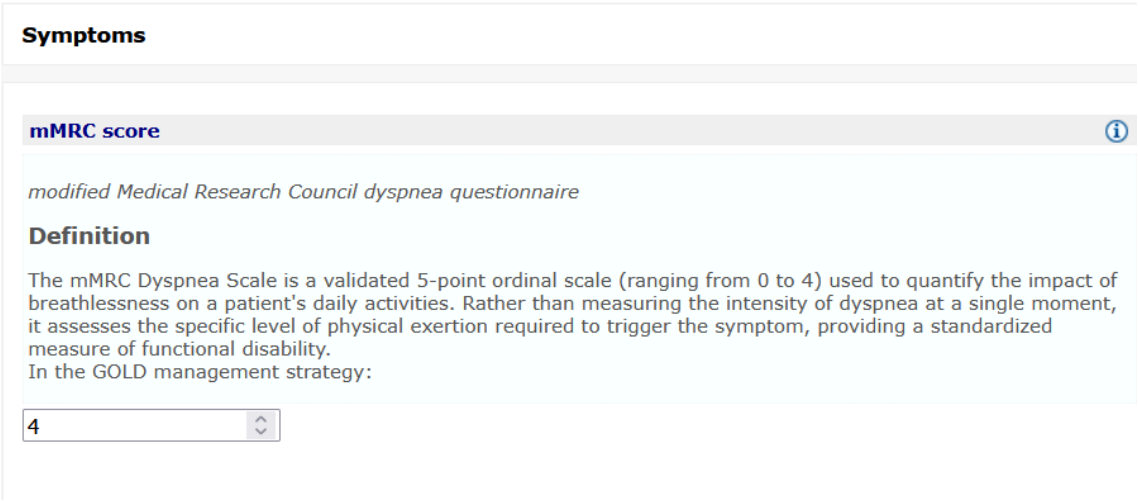


Figura 24: Tester – Definizione della gravità della dispnea.

4. Riferire eventuali effetti collaterali dell'ICS: nessuno (Figura 25)

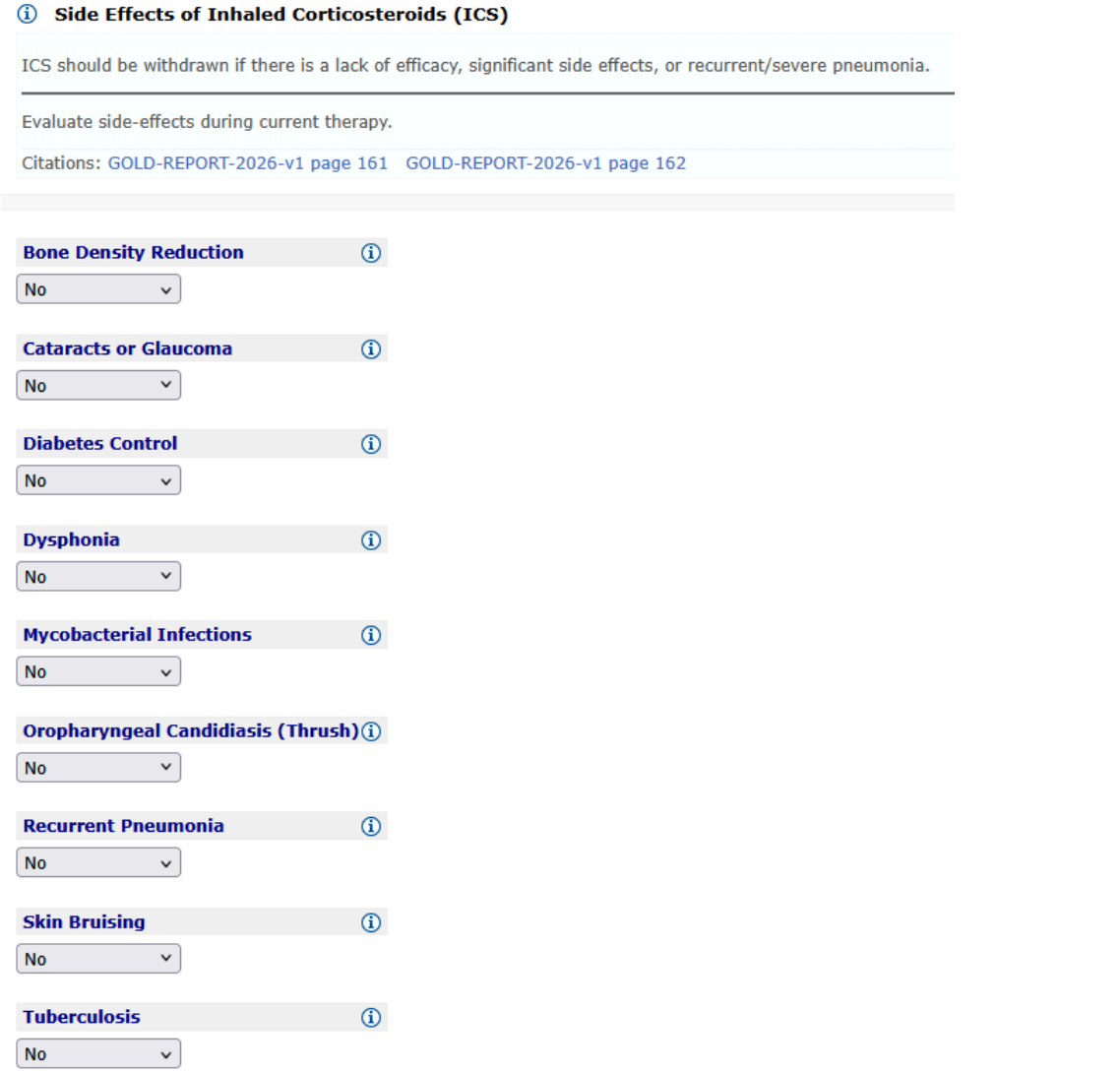


Figura 25: Tester – Eventuali effetti collaterali dell'ICS.

5. Definire conteggio degli eosinofili nel sangue: 48 cellule/µL (Figura 26)

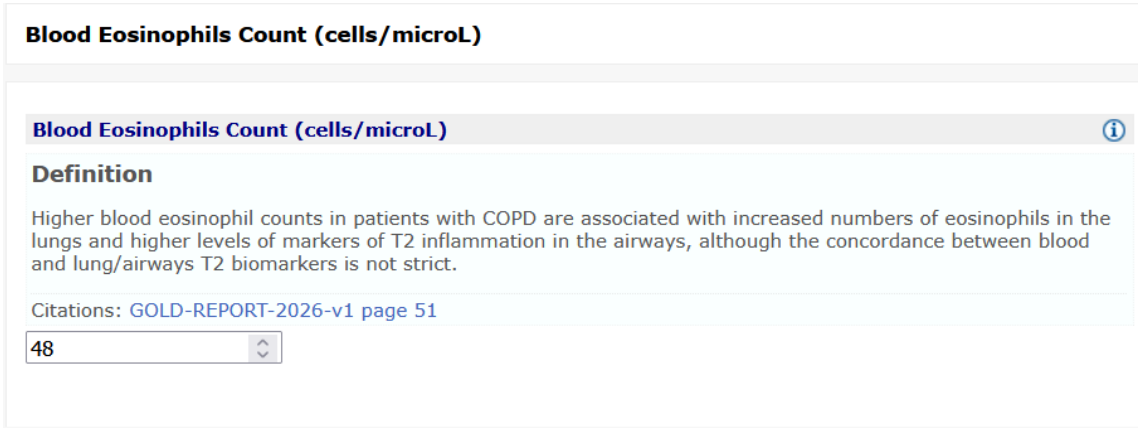


Figura 26: Tester – Riportare il conteggio degli eosinofili.

6. Definire eventuali indicazioni speciali (Ex fumatore, Bronchite cronica, Spirometria): nessuna (Figura 27)

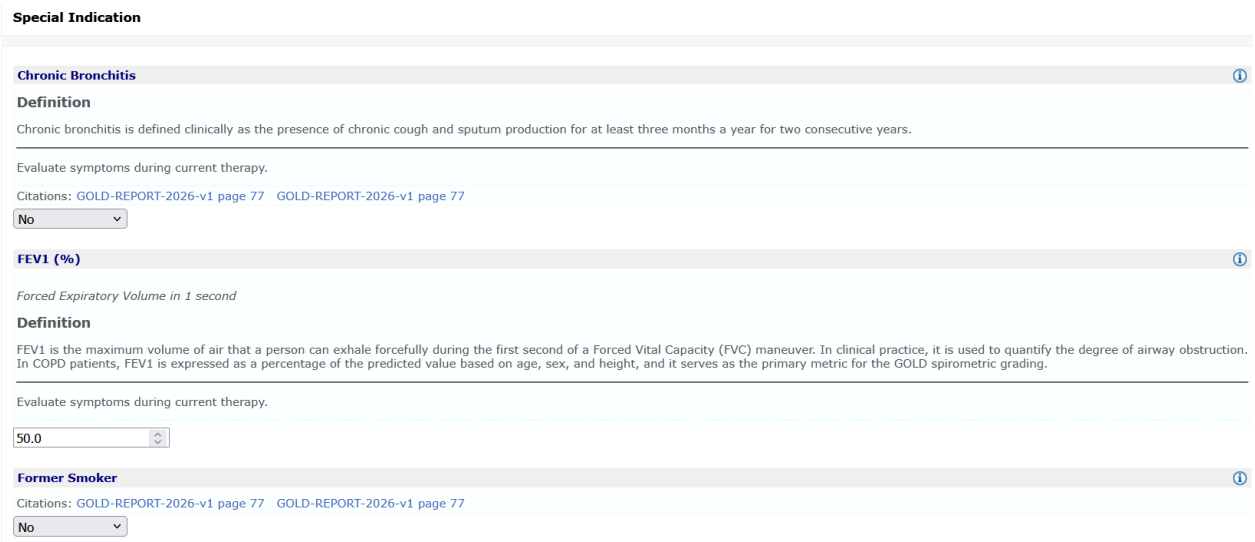


Figura 27: Tester – Riportare indicazioni speciali.

7. Nuova terapia raccomandata: Triplice terapia (LABA+LAMA+ICS) (Figura 28)

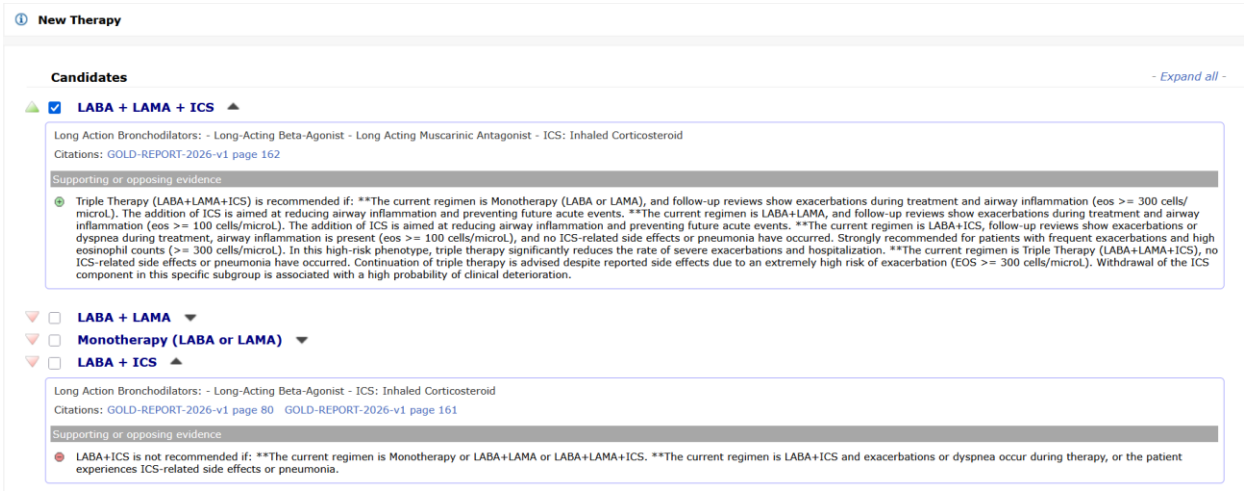


Figura 28: Tester – Selezione della terapia di dimissione

La terapia raccomandata dal sistema al termine del percorso decisionale coincide con la terapia di dimissione effettivamente prescritta al paziente dal personale clinico (*discharge therapy*: LABA+LAMA+ICS), a conferma della coerenza del comportamento del Composer rispetto alla pratica clinica reale documentata nel caso. L'implementazione di questo sistema esperto risulta rilevante ai fini del progetto, in quanto dimostra, attraverso un sistema rigido e deterministico, l'importanza e la validità dei sistemi basati sulla conoscenza nel supporto alla decisione clinica.

L'implementazione di un sistema di questo tipo è indubbiamente più complessa e onerosa rispetto a un'architettura RAG, in termini sia di tempo di sviluppo sia di manutenzione continua della base di conoscenza formalizzata. Tuttavia, è proprio questa manutenzione, interamente affidata a esperti del dominio, a garantire l'affidabilità e la trasparenza del sistema, caratteristiche imprescindibili in un contesto clinico ad alto rischio come quello qui considerato. La validazione condotta sui casi clinici forniti dagli ospedali partner ha permesso di verificare concretamente tale affidabilità, confermando la coerenza delle raccomandazioni prodotte dal sistema con le decisioni terapeutiche effettivamente adottate nella pratica clinica.

Se il *Composer* appena descritto rappresenta la concretizzazione del paradigma dei sistemi esperti nel progetto GUIDE-AI, la parte restante del capitolo è dedicata all'implementazione del secondo paradigma discusso nella tesi: l'architettura RAG. Il punto di partenza di tale implementazione è la costruzione del *Knowledge Graph*, che costituisce l'infrastruttura di rappresentazione della conoscenza clinica su cui si fonda l'intero sistema RAG.

5.5 Knowledge Graph

La terza fase del progetto ha previsto l'implementazione del *Knowledge Graph*, costituito da due tipi di grafi.

Il primo tipo è costituito dai grafi di riferimento, che rappresentano la conoscenza clinica standardizzata: il grafo di SNOMED CT, che descrive i concetti clinici e le relazioni tra essi, e il grafo dell'ATC, che organizza i principi attivi in una gerarchia di classi. A questi si aggiunge DrugClass, un piccolo grafo costruito per questo lavoro, che raccoglie le classi di farmaci citate nella linea guida sulla BPCO (LABA, LAMA, ICS, SABA e SAMA). Questi grafi sono indipendenti dal documento.

Il secondo tipo è il grafo del documento, che rappresenta la struttura della linea guida: la suddivisione in sezioni, sottosezioni, chunk e figure, e l'ordine in cui compaiono.

In questo lavoro sono stati realizzati tre passaggi. (1) La linea guida è stata selezionata e suddivisa in chunk, e le figure convertite in testo (paragrafo 5.5.1). (2) Il documento è stato modellato come grafo (paragrafo 5.5.2). (3) I grafi di riferimento sono stati collegati tra loro. (paragrafo 5.5.3).

Il collegamento tra i due tipi di grafi è invece uno sviluppo futuro: consiste nel mappare il testo dei chunk con i concetti SNOMED CT.

5.5.1 Base della conoscenza

Come base della conoscenza da cui partire per costruire il grafo del documento, è stata utilizzata la linea guida sulla BPCO GOLD 2026 [3]. Un'accurata analisi ha permesso di estrarne unicamente le sezioni più significative ai fini della definizione del percorso terapeutico di mantenimento, mantenendo la struttura originaria in paragrafi e sottoparagrafi. La selezione ha condotto a mantenere i capitoli riportati in Tabella 4.

Titolo capitolo	Pagine di riferimento
Pharmacological maintenance treatment	71-75
Initial pharmacological management	75-76
Follow-up pharmacological management	76-80
Pharmacological treatment	114-116
Maintenance therapy for stable COPD	154-165

Tabella 4: Capitoli della linea guida della BPCO (GOLD 2026) in cui viene discusso il percorso terapeutico

I capitoli in esame presentano delle figure essenziali per la definizione del percorso terapeutico (Tabella 5).

Titolo figura	Pagine di riferimento
Figura 3.7 Diagnosis and Management Cycle	72
Figura 3.8 Initial Pharmacological Treatment	73
Figura 3.9 Follow-up Pharmacological Treatment	74
Figura 3.10 Factors to Consider when Initiating ICS Treatment	78
Figura 3.11 Evidence Supporting Use of Biologic in the Treatment of COPD	79
Figura 3.12 Management of Patients Currently on LABA+ICS	80
Figura A3.1 Maintenance Medications in COPD	156
Figura A3.2 Bronchodilators in Stable COPD	157
Figura A3.3 Anti-inflammatory Maintenance Therapy	160

Tabella 5: Illustrazioni della linea guida della BPCO (GOLD 2026) in cui viene discusso il percorso terapeutico

Le raffigurazioni citate rappresentano *flowchart* di processi terapeutici o tabelle contenenti farmaci, dosaggi e combinazioni. Il loro contenuto sta nella struttura: le condizioni, le diramazioni, le azioni a cui portano. Una semplice descrizione in prosa avrebbe reso questa struttura implicita e ambigua. Si è scelto quindi un formato strutturato e testuale, l'XML, per tre motivi. In primo luogo, esprime in modo esplicito le gerarchie e le relazioni (ad esempio una condizione, i suoi rami

e l'azione che ne deriva) tramite elementi annidati e attributi, preservando la logica decisionale del diagramma originale. In secondo luogo, è un formato testuale: ogni figura può essere memorizzata come nodo nel grafo, come per il testo della linea guida. Infine, è verificabile: la struttura può essere confrontata visivamente con la figura originale, cosa non possibile con una descrizione discorsiva.

La traduzione delle immagini in linguaggio XML è stata eseguita fornendo all'LLM Claude (*versione: Claude Sonnet 5*) le immagini di interesse, chiedendone una versione testuale in formato XML. I risultati sono stati successivamente analizzati e validati manualmente, con l'ausilio del software Obsidian (*versione: 1.12.4*, un'applicazione di gestione della conoscenza personale, basata sulla scrittura e l'organizzazione di note in formato Markdown, salvate localmente come file di testo)[96], che riceve in input il testo XML e ne mostra la struttura grafica corrispondente. Questo doppio controllo consente di verificare la validità del testo generato, individuare eventuali errori e correggerli di conseguenza. Va infine notato che tutti gli XML prodotti dal modello Claude presentavano metadati di scarsa utilità informativa, che sono stati pertanto scartati. Ad esempio, in un documento XML relativo alla Figura 3.7 compariva l'elemento:

```
<step id="initiate_treatment" number="1" label="Initiate Treatment"
style="green">
    <text>Proceed to Figure 3.8</text>
    <subtitle>INITIAL Pharmacological Treatment</subtitle>
</step>
```

in cui la precisazione del colore verde risulta superflua ai fini della rappresentazione della conoscenza.

Per chiarire i passaggi sopra citati, si presenta un esempio concreto, relativo alla Figura 3.12 - *Management of Patients Currently on LABA+ICS*. (Figura 29)

1. Estrazione dell'immagine dal testo delle linee guida (Figura 29)

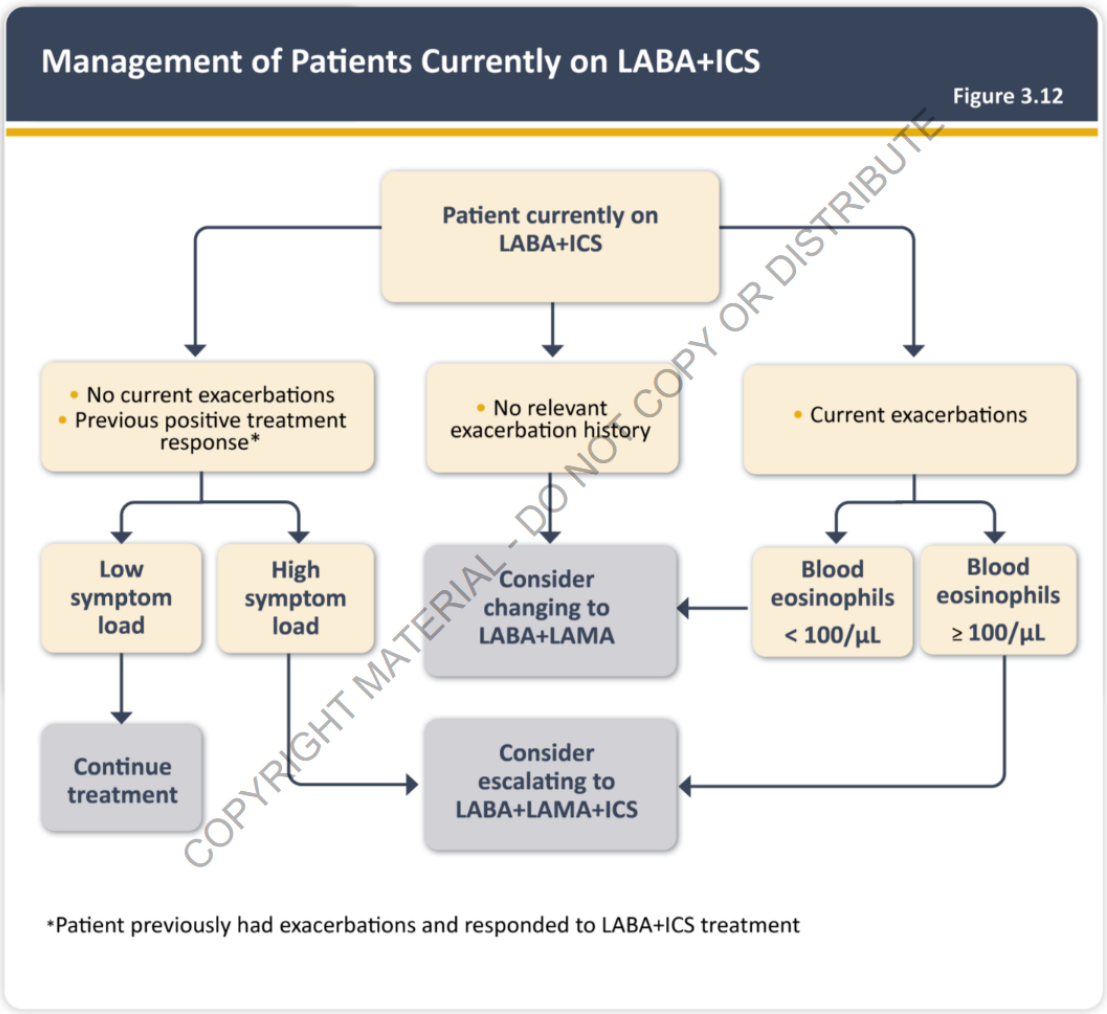


Figura 29: grafico che orienta la decisione medica riguardo alla gestione dei pazienti attualmente in cura con LABA+ICS.

2. Conversione in XML, tramite il modello Claude (prompt all'Appendice 1):

```
<?xml version="1.0" encoding="UTF-8"?>
<diagram title="Management of Patients Currently on LABA+ICS"
figure="3.12">

  <start id="root" label="Patient currently on LABA+ICS"/>

  <branches from="root">

    <branch id="left" condition="No current exacerbations AND
Previous positive treatment response*">
      <sub_branches from="left">
        <branch id="left_low" condition="Low symptom load">
          <outcome id="continue" action="Continue treatment"
style="gray"/>
        </branch>
        <branch id="left_high" condition="High symptom load">
          <transition to="escalate"/>
        </branch>
      </sub_branches>
    </branch>

    <branch id="center" condition="No relevant exacerbation
history">
      <transition to="change_to_laba_lama"/>
    </branch>

    <branch id="right" condition="Current exacerbations">
      <sub_branches from="right">
        <branch id="right_low_eos" condition="Blood eosinophils
&lt; 100/µL">
          <transition to="change_to_laba_lama"/>
        </branch>
        <branch id="right_high_eos" condition="Blood eosinophils ≥
100/µL">
          <transition to="escalate"/>
        </branch>
      </sub_branches>
    </branch>

  </branches>

  <actions>
    <action id="change_to_laba_lama" label="Consider changing to
LABA+LAMA" style="gray"/>
    <action id="escalate" label="Consider escalating to
LABA+LAMA+ICS" style="gray"/>
  </actions>

  <footnotes>
    <footnote marker="*">Patient previously had exacerbations and
responded to LABA+ICS treatment.</footnote>
  </footnotes>

</diagram>
```

3. Validazione tramite Obsidian, che traduce l'XML nella sua rappresentazione grafica corrispondente (Figura 30).

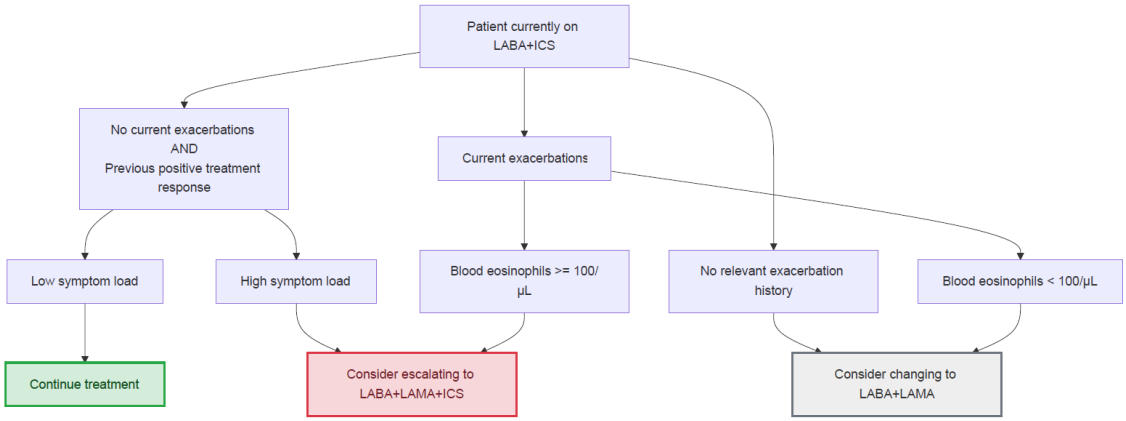


Figura 30: grafico, creato con Obsidian da documento XML, che orienta la decisione medica riguardo alla gestione dei pazienti attualmente in cura con LABA+ICS.

4. Ripulitura manuale dai metadati non informativi (nell'esempio, gli attributi style presenti in più punti del documento sarebbero da rimuovere in fase di pulizia, essendo privi di valore semantico):

```
<?xml version="1.0" encoding="UTF-8"?>
<diagram title="Management of Patients Currently on LABA+ICS"
figure="3.12">

  <start id="root" label="Patient currently on LABA+ICS"/>

  <branches from="root">

    <branch id="left" condition="No current exacerbations AND
Previous positive treatment response*">
      <sub_branches from="left">
        <branch id="left_low" condition="Low symptom load">
          <outcome id="continue" action="Continue treatment"
style="gray"/>
        </branch>
        <branch id="left_high" condition="High symptom load">
          <transition to="escalate"/>
        </branch>
      </sub_branches>
    </branch>

    <branch id="center" condition="No relevant exacerbation
history">
      <transition to="change_to_laba_lama"/>
    </branch>

    <branch id="right" condition="Current exacerbations">
```

```
      <sub_branches from="right">
        <branch id="right_low_eos" condition="Blood eosinophils
&lt; 100/µL">
          <transition to="change_to_laba_lama"/>
        </branch>
        <branch id="right_high_eos" condition="Blood eosinophils ≥
100/µL">
          <transition to="escalate"/>
        </branch>
      </sub_branches>
    </branch>

  </branches>

  <actions>
    <action id="change_to_laba_lama" label="Consider changing to
LABA+LAMA" style="gray"/>
    <action id="escalate" label="Consider escalating to
LABA+LAMA+ICS" style="gray"/>
  </actions>

  <footnotes>
    <footnote marker="*">Patient previously had exacerbations and
responded to LABA+ICS treatment.</footnote>
  </footnotes>

</diagram>
```

Questo processo di selezione, estrazione e validazione ha permesso di creare una base di conoscenza solida e scientificamente fondata, che ha costituito il nucleo informativo del *Knowledge Graph*.

5.5.2 Modellazione delle Linee Guida nel database a grafo

Per la costruzione del *Knowledge Graph*, i documenti selezionati sono stati organizzati e processati mantenendo la loro struttura originaria: i testi sono stati suddivisi in *chunk* corrispondenti a paragrafi e sottoparagrafi, così da preservare la coerenza semantica e facilitare il successivo recupero delle informazioni; allo stesso modo, anche ciascuna figura convertita in formato XML costituisce un *chunk* a sé stante.

Si passa, così, alla fase di mappatura, fondamentale per identificare come capitoli e figure interagiscano tra loro. Tale struttura costituirà la base dell'organizzazione del *Knowledge Graph*. Come mostrato in Figura 31, sussiste un rapporto di sequenzialità tra i capitoli iniziali *Pharmacological maintenance treatment*, *Initial pharmacological management* e *Follow-up pharmacological management*; questi

ultimi due sono a loro volta collegati ai due capitoli che trattano nello specifico i farmaci: *Pharmacological treatment* e *Maintenance therapy for stable COPD*. Infine, le figure sono state collegate ai capitoli in cui vengono citate.

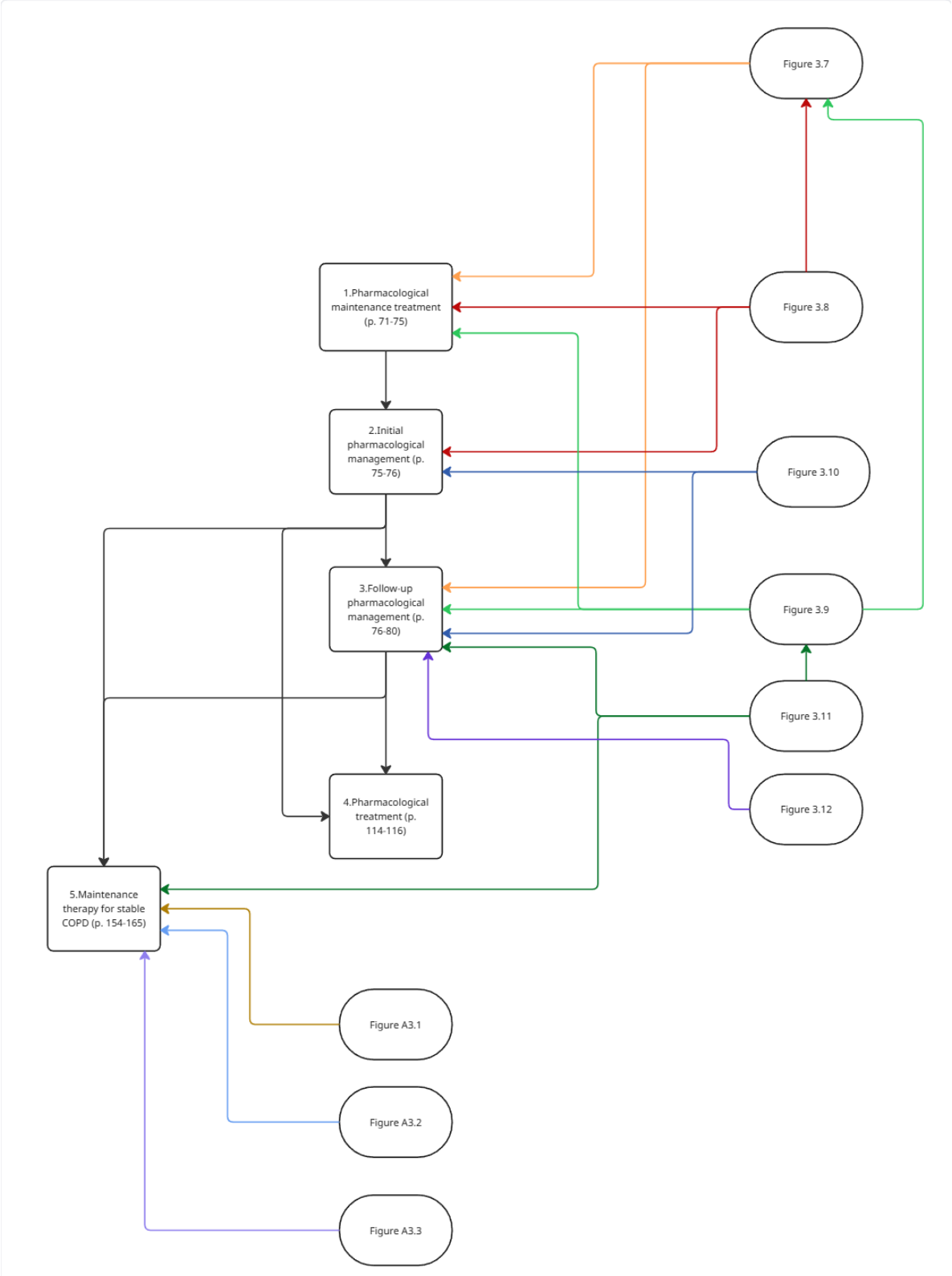


Figura 31: mappatura di capitoli e immagini selezionati dalla linea guida della BPCO.

Ogni entità individuata è stata rappresentata come nodo dotato di proprietà descrittive (es. testo originale, identificativi), mentre le relazioni tra le entità sono state modellate come archi etichettati. Questo approccio ha permesso di costruire una rappresentazione coerente e navigabile delle informazioni contenute nelle linee guida.

La Tabella 6 riassume le entità usate per modellizzare le linee guida. La strategia adottata rispecchia la scelta di rappresentare nel grafo la gerarchia dei paragrafi contenuti nei documenti.

Nome	Definizione	Attributi
Chunk	Parte di testo ottenuta con il chunking, corrisponde ad un paragrafo o sottoparagrafo della linea guida.	**chunkId**: codice identificativo del chunk **title**: titolo **parentId**: codice identificativo del paragrafo di appartenenza **text**: testo del paragrafo
Guideline	Intera linea guida	**guidelineId**: ID della linea guida **title**: titolo della linea guida
Section	Paragrafo	**sectionId**: codice identificativo del paragrafo **title**: titolo **numChunks**: numero di chunk contenuti **numSubSections**: numero di sottoparagrafi che lo compongono
Subsection	Sottoparagrafo	**subSectionId**: codice identificativo del sottoparagrafo **title**: titolo **sectionName**: nome del paragrafo di appartenenza **numChunks**: numero di chunk contenuti
Figure	Figura in formato XML	**figureId**: codice identificativo della figura **title**: titolo **text**: testo in formato XML che descrive l'immagine

Tabella 6: Tipologie di entità utilizzate per modellizzare il KG delle linee guida

Definite le entità, sono state introdotte le relazioni che le collegano (Figura 32). La relazione HAS_SECTION lega Guideline a Section, e HAS_SUBSECTION lega Section a Subsection: insieme rendono esplicita la gerarchia interna del documento. Poiché sia Section che Subsection contengono porzioni di testo, entrambe sono collegate ai rispettivi chunk tramite la relazione HAS_CHUNK. Tra entità dello stesso tipo (Chunk, Section e Subsection) sono presenti archi NEXT, che preservano l'ordine sequenziale del testo.

Per finire, i nodi Chunk e Figure possono rimandare ad altre figure per mezzo della relazione REFERENCES_FIGURE.

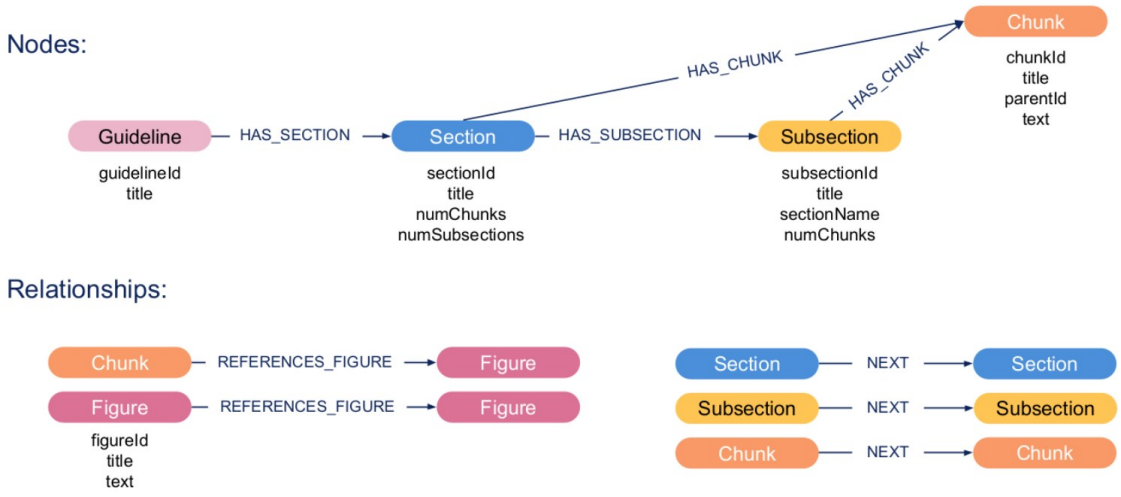


Figura 32: nodi e relazioni che costituiscono il Knowledge Graph.

Il KG risultante è riportato in Figura 33.

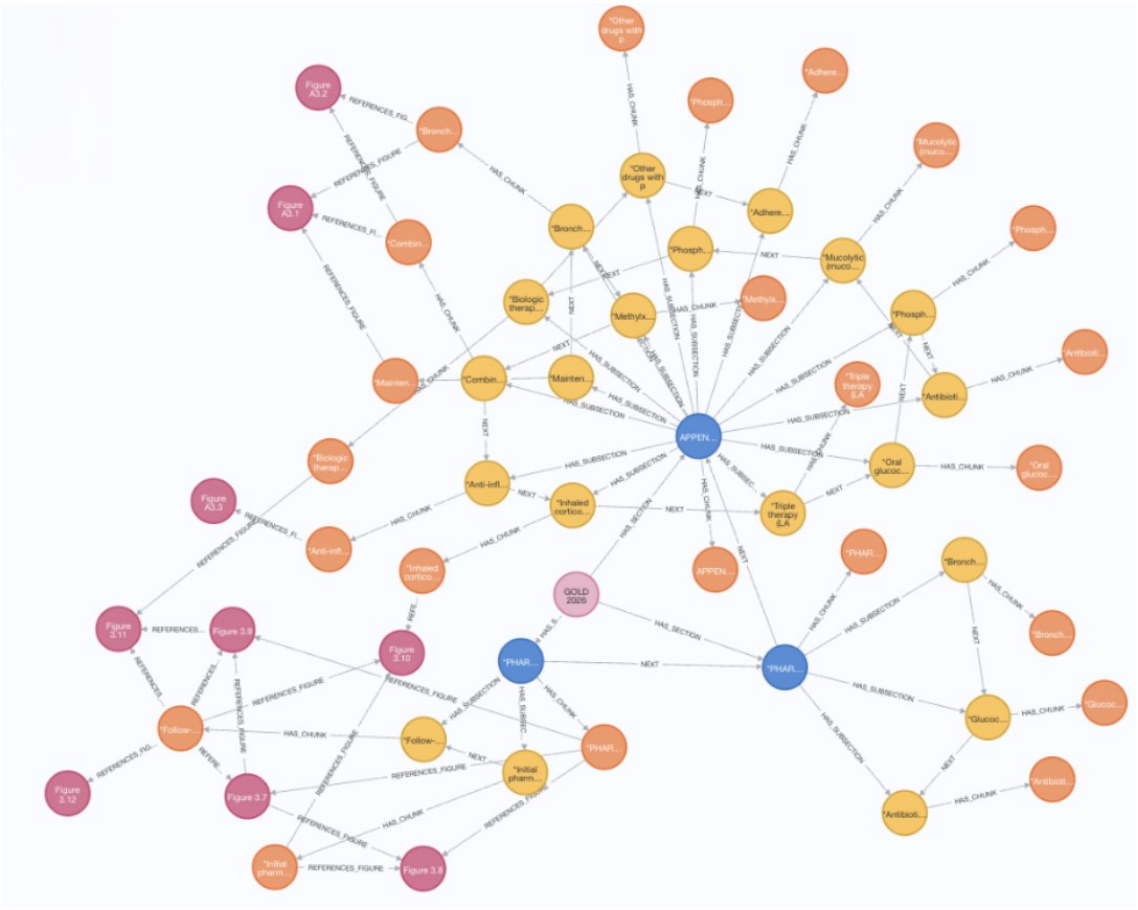


Figura 33: grafo di conoscenza delle linee guida della BPCO.

Questa strutturazione consente di navigare il contenuto testuale in modo sia sequenziale, seguendo l'ordine dei paragrafi, sia gerarchico, preservando così la coerenza semantica e il contesto informativo di partenza.

5.5.3 Integrazione con SNOMED CT e ATC

Il grafo che modella il contenuto delle linee guida è affiancato da tre grafi che contengono la conoscenza clinica:

- SNOMED CT: terminologia clinica standardizzata;
- ATC: il sistema internazionale di classificazione anatomico-terapeutico-chimica dei farmaci;

- DrugClass: un sistema di classificazione generale dei farmaci, costruito specificatamente per la patologia in esame.

Quest'ultimo è stato creato perché le linee guida della BPCO raccomandano classi generali di farmaci, anziché principi attivi specifici:

- LABA: agonisti beta-2 a lunga durata d'azione (*Long-Acting Beta-2 Agonist*);
- LAMA: antimuscarinici a lunga durata d'azione (*Long-Acting Muscarinic Antagonist);*
- ICS: corticosteroidi per via inalatoria (*Inhaled Corticosteroids*);
- SABA: agonisti beta-2 a breve durata d'azione (*Short-Acting Beta-2 Agonist*);
- SAMA: antimuscarinici a breve durata d'azione (*Short-Acting Muscarinic Antagonist*).

Per ricondursi alla linea guida in fase di interrogazione (*querying*) è dunque necessario ricollegare i principi attivi formalizzati nell'ATC alla classe farmacologica di appartenenza, così da poterli collegare ai *chunk* in cui tale classe viene discussa.

La formalizzazione dei concetti prevede la presenza di tre grafi, le cui dimensioni sono riportate in Tabella 7.

Grafo di conoscenza	# Nodi
SNOMED CT	378.290
ATC	7.024
DrugClass	5

Tabella 7: Numero di nodi per ciascun grafo di conoscenza.

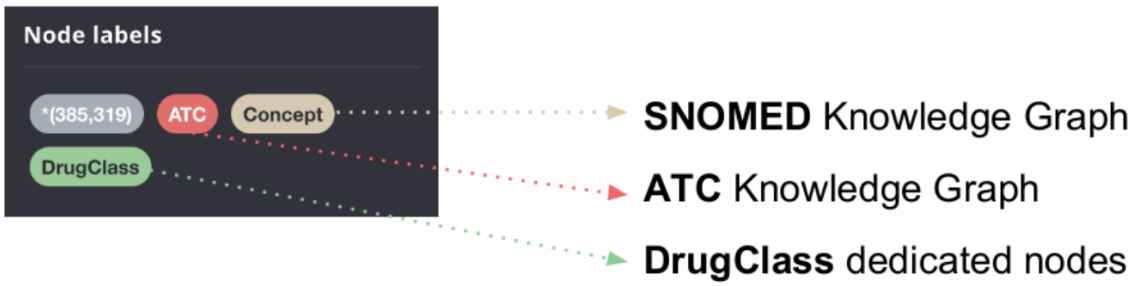


Figura 34: etichette dei nodi dei grafi di SNOMED, ATC, DrugClass.

Come mostra la Figura 34, i nodi del grafo di SNOMED CT sono etichettati come Concept, mentre i nodi dei grafi ATC e DrugClass sono etichettati in maniera omonima al nome del grafo di appartenenza. La Tabella 8 riassume gli attributi specifici delle entità:

Nome	Attributi
Concept	**id**: codice identificativo SNOMED CT **descriptions**: definizioni e varianti linguistiche che descrivono il termine clinico
ATC	**cui**: codice unico del concetto nel sistema metathesaurus UMLS (Unified Medical Language System) **id**: codice standard ATC **level**: livello gerarchico nella struttura ad albero dell'ATC **name**: descrizione testuale ufficiale associata al codice ATC **obsolete**: valore booleano per indicare se il nodo è ancora attivo nel vocabolario corrente
DrugClass	**ClassName**: nome della classe farmacologica

Tabella 8: Attributi specifici delle entità dei grafi SNOMED CT, ATC e DrugClass.

I tre grafi sono uniti tra loro tramite delle relazioni. Il collegamento tra SNOMED CT e ATC è stato definito tramite le API REST dell'UMLS. (Paragrafo 3.3.3) Per ciascun concetto SNOMED CT corrispondente a un principio attivo si è ricavato

il CUI e se ne sono cercati i codici ATC associati. Uno stesso concetto può corrispondere a più codici ATC. I nodi Concept sono quindi collegati ai nodi ATC tramite la relazione HAS_ATC_CODE, e il CUI è conservato come attributo dei nodi ATC (Tabella 8). A loro volta, i termini ATC possono identificare farmaci appartenenti alle classi citate nelle linee guida e sono pertanto collegati ai nodi DrugClass tramite la relazione HAS_DRUG_CLASS (Figura 35).

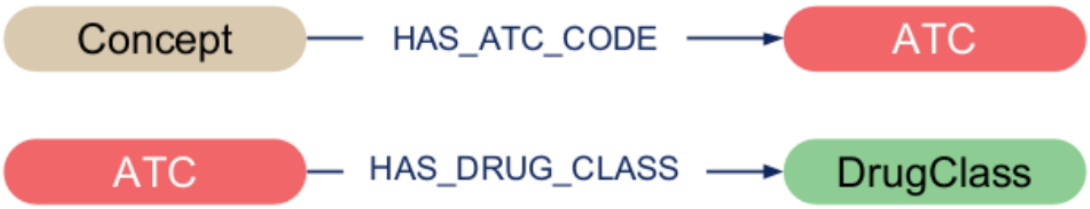


Figura 35: relazione SNOMED-ATC (sopra); relazione ATC-DrugClass (sotto)

Le figure seguenti illustrano alcuni esempi di sottografi risultanti:

- Figura 36: esempio di relazioni tra i nodi ATC e DrugClass;
- Figura 37: sottografo risultante a partire dai nodi DrugClass;
- Figura 38: sottografo risultante a partire da un principio attivo.

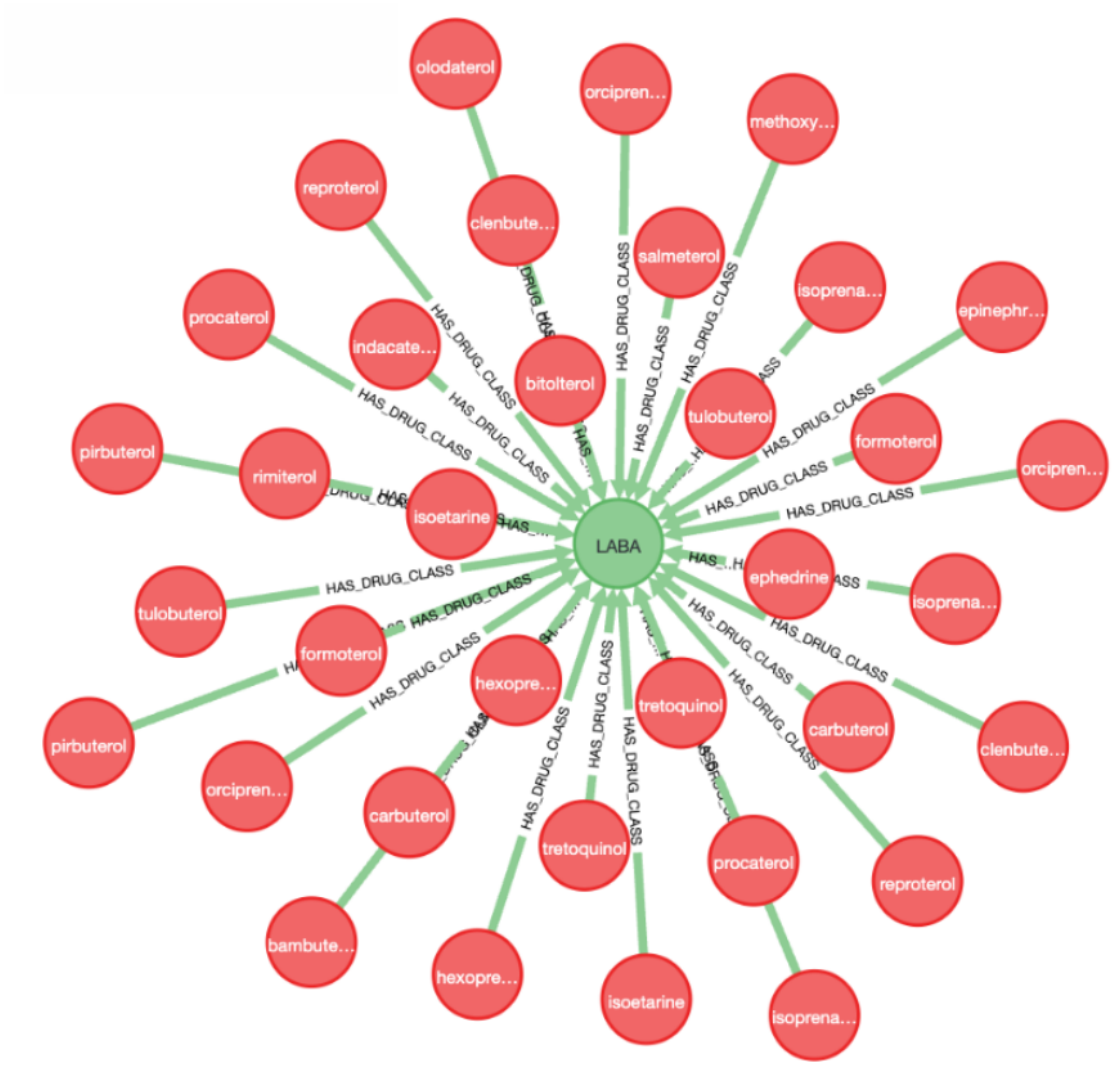


Figura 36: Termini ATC appartenenti alla classe farmacologica LABA.

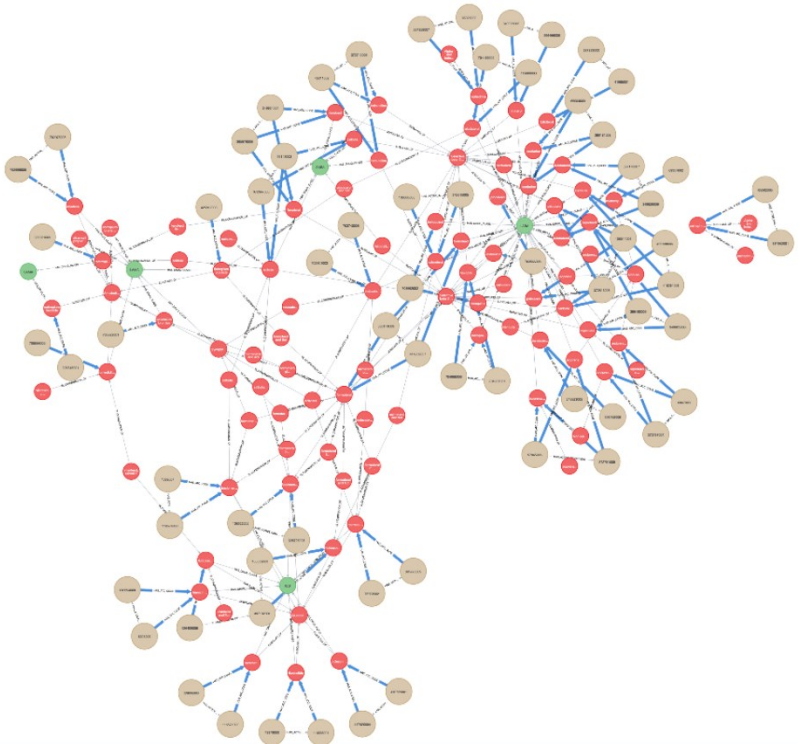


Figura 37: sottosezione del grafo risultante partendo da i 5 nodi DrugClass, i nodi ATC ad essi associati fino a 2 gradi di distanza e i nodi SNOMED legati agli ATC ad un grado di distanza.

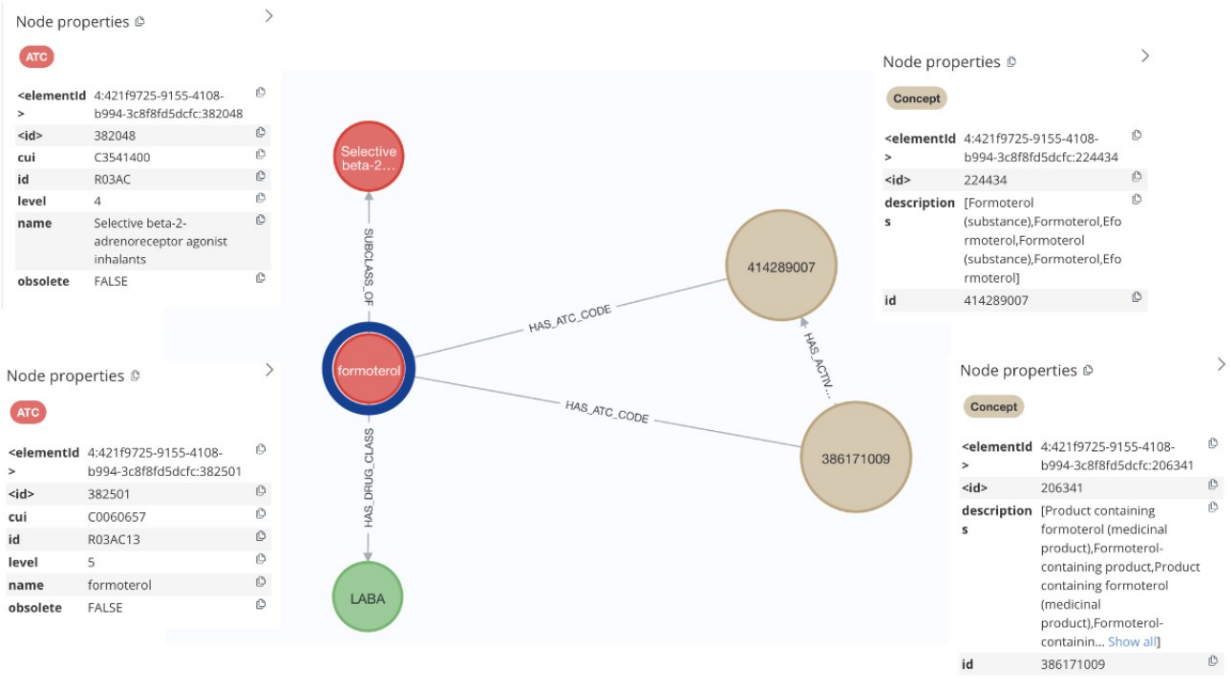


Figura 38: nodo ATC riferito al formoterolo e tutti i nodi del grafo a cui è collegato entro un grado di distanza.

Uno sviluppo futuro prevederà il collegamento dei *chunk* ai concetti SNOMED CT contenuti in essi. Per l'estrazione di tali termini, al fine di rendere il processo più scalabile ed efficiente, è stato applicato un sistema dedicato di estrazione automatica dei concetti clinici pertinenti (*concept extractor*), basato sull'integrazione di modelli linguistici.

5.6 Estrattore dei concetti SNOMED CT

L'ultima fase del progetto di tesi si è incentrata sull'estrazione dei concetti SNOMED CT dai frammenti di linea guida, applicando il processo descritto nel capitolo 4, sviluppato in precedenza per un progetto affine relativo alla diagnosi dell'artrite reumatoide. Il contributo specifico di questo lavoro è consistito nell'adattamento di tale strumento al dominio della BPCO e alla linea guida qui analizzata.

5.6.1 Costruzione dei Dizionari

Il primo passaggio ha previsto la costruzione di dizionari terminologici adeguati su cui eseguire la ricerca basata su dizionario (descritta nel paragrafo 4.2), poiché quello impiegato nel progetto originario conteneva termini SNOMED CT specifici del dominio dell'artrite reumatoide.

In primo luogo, è stato costruito un dizionario contenente tutti i concetti SNOMED CT attualmente attivi. A tal fine, è stato scaricato dal portale di distribuzione di SNOMED International il file delle descrizioni dell'edizione internazionale di SNOMED CT, nella release del 1° giugno 2026. Il file è in formato RF2 (*Release Format 2*), il formato di distribuzione standard di SNOMED CT: un file di testo tabulare in cui ogni riga è una descrizione, cioè un termine in lingua inglese con cui viene designato un concetto. Per ogni concetto sono presenti il nome completo e i sinonimi. Ogni riga riporta inoltre lo stato della descrizione (attiva o inattiva) e la data di rilascio della versione. Il file è nella versione *Full*, che conserva lo storico di tutte le modifiche. Per ciascuna descrizione attiva sono stati estratti il termine e l'identificativo del concetto a cui si riferisce, salvati in un file CSV.

Delle 2.931.190 righe di descrizione contenute all'interno del dataset SNOMED CT, 2.616.876 sono state scritte nel file CSV, poiché 314.314 sono state escluse in quanto non attive.

Poiché più righe di descrizione facevano riferimento a coppie termine-id identiche, i duplicati sono stati rimossi mantenendo la prima occorrenza. Al termine dell'operazione, le voci uniche mantenute risultano 1.624.924, a fronte di 991.952 duplicati rimossi.

Successivamente è stato creato un secondo dizionario più mirato, ottenuto filtrando il dizionario completo tramite il *Knowledge Graph* descritto nel paragrafo 5.5: a partire dal concetto di *Chronic obstructive lung disease* (*disorder*) (codice: 13645005), sono stati selezionati tutti i concetti a esso collegati entro una distanza massima di tre gradi di separazione (ovvero di 3 relazioni) all'interno del grafo. Il dizionario risultante contiene 57.330 voci di descrizione. La creazione di un dizionario più contenuto e specifico per la patologia analizzata ha l'obiettivo di ridurre il tasso di falsi positivi (definiti al paragrafo 4.5), circoscrivendo la ricerca basata su dizionario a un dizionario più contenuto e specifico per la patologia analizzata.

5.6.2 Creazione del file di validazione

Per valutare le prestazioni dell'estrattore di concetti (paragrafo 5.6.3) occorre confrontare i concetti che estrae con un insieme di riferimento, la *ground truth*, che in questo lavoro è raccolta in un file di validazione. La qualità di questo riferimento determina l'affidabilità delle metriche di valutazione (*precision*, *recall* e F1-score): una *ground truth* incompleta o imprecisa porterebbe a una stima distorta delle prestazioni. Generalmente questo compito viene affidato a clinici esperti, che analizzano la linea guida e, per ciascuna sezione di interesse, individuano i concetti clinici rilevanti e il corrispondente concetto SNOMED CT.

In questo caso, tale compito è stato svolto direttamente nell'ambito del presente lavoro. I concetti clinici sono stati estratti manualmente e successivamente si è

quindi scelto di supportare l'individuazione dei concetti clinici tramite *Amazon Comprehend Medical* [97], un servizio di NLP offerto da Amazon Web Services (AWS), basato su modelli di *machine learning* addestrati per estrarre informazioni sanitarie da testi medici non strutturati. Si tratta di un servizio completamente gestito: non richiede la costruzione o l'addestramento di alcun modello da parte dell'utente, ed è accessibile tramite API. [98]

Tra le operazioni offerte, quella impiegata in questo lavoro è *InferSNOMEDCT*, che opera in due passaggi. In primo luogo, individua nel testo le entità mediche e le assegna a una categoria (condizioni mediche, anatomia, test, trattamenti e procedure). Poi collega ciascuna entità ai cinque concetti più probabili della terminologia SNOMED CT. Per ogni entità rilevata, il servizio restituisce il testo dell'entità, la categoria, la posizione di inizio e fine nel testo e un punteggio di confidenza sul riconoscimento. Tale valore oscilla tra 0 e 1 ed esprime il grado di fiducia del modello nella propria predizione. Poiché la modalità di calcolo non è documentata pubblicamente, tale punteggio è stato interpretato come un indicatore relativo dell'ordine di preferenza del modello, e non come una probabilità calibrata di correttezza.

L'output viene restituito in formato JSON ed è disponibile per testi in inglese. La restituzione dei cinque concetti più probabili, anziché del solo migliore, ha reso lo strumento adatto a supportare la costruzione della *ground truth*.

Poiché sarebbe metodologicamente scorretto validare un estrattore di concetti basato su intelligenza artificiale utilizzando un altro strumento anch'esso soggetto ad allucinazioni ed errori, si è ritenuto essenziale analizzare manualmente e validare, vocabolo per vocabolo, l'intera *ground truth*. È stata quindi condotta un'individuazione manuale dei concetti clinici presenti nella linea guida riguardante la BPCO, sulla base di questi sono stati validati i concetti prodotti in output da *Comprehend Medical*, eliminando i termini errati e integrando quelli mancanti. Sono stati considerati errati:

1. termini troppo lontani semanticamente dal significato del concetto estratto dal testo (es. per il termine "*Exacerbation*" vengono estratti termini che riportino questa parola, ma che si riferiscano a riacutizzazioni di altre malattie, come "*Exacerbation of osteoarthritis* (*disorder*)");
2. termini scarsamente significativi se non associati al contesto (prendendo sempre ad esempio il termine "*Exacerbation*", il concetto "*Exacerbated* (*qualifier value*)" risulta troppo vago ed associabile ad una qualunque forma di riacutizzazione);
3. termini associati erroneamente ad acronimi (es. l'acronimo "CRP" riferito alla Proteina C-Reattiva, è stato erroneamente associato a "*Comprehensive metabolic panel* (*procedure*)", "*Complex lung function test* (*procedure*)", "*CR type 2 receptor measurement* (*procedure*)", "*Cytidine monophosphate-N-acetylneuraminate monooxygenase* (*substance*)", tutti concetti scartati di conseguenza).

Una volta completata questa operazione, sono stati estratti i concetti unici dal file JSON restituito da *Comprehend Medical* ed è stato creato un file Excel contenente, nella prima colonna, il testo del *chunk* di riferimento, nella seconda i termini SNOMED CT corrispondenti e nella terza i relativi codici SNOMED CT, mantenendo lo stesso ordine dei termini nella colonna precedente. Una volta costruita la *ground truth*, essa è stata convertita in formato JSON, per garantire che tutte le informazioni vengano preservate senza perdita. Un file JSON è, infatti, un formato di testo leggero e leggibile, ampiamente utilizzato per il trasporto e lo scambio di informazioni tra sistemi diversi, come un server e un'applicazione web. Inoltre, un file JSON può contenere un'enorme quantità di informazioni, poiché non possiede limiti teorici di dimensione strutturale, se non quelli fisici della memoria del computer o del software che lo elabora. [99]

Si illustrano di seguito i passaggi seguiti per la costruzione della *ground truth*, con i relativi esempi.

1. Analisi della linea guida e individuazione dei concetti clinici presenti. (Figura 39)

Maintenance therapy for stable COPD

Pharmacotherapy for COPD is currently focused on symptoms and exacerbations. FEV1 decline has been considered a surrogate for the natural course of the disease. In this context, studies have been performed to evaluate if pharmacotherapy may have an impact on the change of FEV1 over time. Individual clinical trials have not been sufficiently conclusive to show that pharmacotherapy can reduce the rate of FEV1 decline. However, a systematic review combining data from nine studies demonstrated a reduction in the rate of FEV1 decline of 5.0 mL/year in active treatment arms compared with placebo arms. The difference between long-acting bronchodilator containing treatment arms and placebo arms was 4.9 mL/year. The difference between ICS containing treatment arms and placebo arms was 7.3 mL/year. Although we need to be aware of the potential benefit of pharmacotherapy in reducing the rate of lung function decline, further research is needed to know which patients are likely to benefit.

The classes of medications commonly used to treat COPD are shown in Figure A3.1. The choice within each class depends on the availability and cost of medication, and the clinical response balanced against side effects. Each treatment regimen needs to be individualized as the relationship between severity of symptoms, airflow obstruction, and severity of exacerbations can differ between patients. The WHO has defined a minimum set of interventions for the management of stable COPD in primary care.

Figura 39: termini clinici individuati all'interno del chunk riferito al capitolo "Maintenance therapy for stable COPD"

2. Analisi delle estrazioni prodotte da *Comprehend Medical*, con eliminazione dei concetti errati e aggiunta dei concetti mancanti.

   Esempio dei concetti estratti utilizzando *Comprehend Medical* per il termine "*Exacerbations*" prima della validazione manuale:


```
{
  "Id": 39,
  "Text": "exacerbations",
  "Category": "MEDICAL_CONDITION",
  "Type": "DX_NAME",
  "Score": 0.5330596566200256,
  "BeginOffset": 1474,
  "EndOffset": 1487,
  "Attributes": [],
  "Traits": [],
  "SNOMEDCTConcepts": [
            {
                "Description": "Exacerbated (qualifier value)",
                "Code": "410661009",
                "Score": 0.04016994312405586
            },
            {
                "Description": "Exacerbation of osteoarthritis
(disorder)",
                "Code": "394991004",
                "Score": 0.013455926440656185
            },
            {
                "Description": "Acute exacerbation of chronic
obstructive airways disease (disorder)",
                "Code": "195951007",
                "Score": 0.012925047427415848
            },
            {
                "Description": "Acute exacerbation of chronic
asthmatic bronchitis (disorder)",
                "Code": "442025000",
                "Score": 0.009820753708481789
            },
            {
                "Description": "Acute infective exacerbation of
chronic obstructive airways disease (disorder)",
                "Code": "285381006",
                "Score": 0.009133712388575077
            }
 ]
}
```

Esempio dei concetti estratti utilizzando *Comprehend Medical* e confermati tramite validazione manuale per il termine "*Exacerbation*s".


```
{
  "Id": 39,
  "Text": "exacerbations",
  "Category": "MEDICAL_CONDITION",
  "Type": "DX_NAME",
  "Score": 0.5330596566200256,
  "BeginOffset": 1474,
  "EndOffset": 1487,
  "Attributes": [],
  "Traits": [],
  "SNOMEDCTConcepts": [
      {
     "Description": "Acute exacerbation of chronic obstructive
airways disease (disorder)",
          "Code": "195951007",
          "Score": 0.012925047427415848
      },
      {
          "Description": "Acute exacerbation of chronic
asthmatic bronchitis (disorder)",
          "Code": "442025000",
          "Score": 0.009820753708481789
      },
      {
          "Description": "Acute infective exacerbation of
chronic obstructive airways disease (disorder)",
          "Code": "285381006",
          "Score": 0.009133712388575077
      }
  ]
}
```

3. Creazione del file di *ground truth* in formato Excel. (Esempio in Tabella 9)

testo	nome_inglese	concept_code
Anti-inflammatory agents To date, exacerbations (e.g., exacerbation rate, patients with at least one exacerbation, time-to-first exacerbation) represent the main clinically relevant endpoint used for efficacy assessment of drugs with anti-inflammatory effects (<imageA3_3.png>).	['Non-steroidal anti-inflammatory agent therapy (procedure)','Non-steroidal anti-inflammatory agent (product)','Non-steroidal anti-inflammatory agent (substance)','Medicinal product acting as antiinflammatory agent (product)','Non-steroidal anti-inflammatory drugs contraindicated (situation)','Acute exacerbation of chronic asthmatic bronchitis (disorder)','Acute exacerbation of chronic bronchitis (disorder)','Acute exacerbation of chronic obstructive bronchitis (disorder)','Acute exacerbation of chronic obstructive airways disease (disorder)']	[713443004, 16403005, 372665008, 330901000, 416759002, 442025000, 425748003, 293241000119100, 195951007]

Tabella 9: riga della ground truth creata in formato Excel riferita al chunk intitolato "Anti-inflammatory agents"

4. Conversione del file Excel in formato JSON. (Esempio alla Figura 40)


```
{
    "testo":"*Anti-inflammatory agents* To date, exacerbations (e.g., exacerbation rate, patients with at least one exacerbation, time-to-first exacerbation) represent the main clinically relevant endpoint used for efficacy assessment of drugs with anti-inflammatory effects (<imageA3_3.png>).",

    "nome_inglese":"['Non-steroidal anti-inflammatory agent therapy (procedure)','Non-steroidal anti-inflammatory agent (product)','Non-steroidal anti-inflammatory agent (substance)','Medicinal product acting as antiinflammatory agent (product)','Non-steroidal anti-inflammatory drugs contraindicated (situation)','Acute exacerbation of chronic asthmatic bronchitis (disorder)','Acute exacerbation of chronic bronchitis (disorder)','Acute exacerbation of chronic obstructive bronchitis (disorder)','Acute exacerbation of chronic obstructive airways disease (disorder)']",

"concept_code":"[713443004,16403005,372665008,330901000,416759002,442025000,425748003,293241000119100,195951007]"
}
```


Figura 40: elemento riferito al chunk intitolato "Anti-inflammatory agents" del file JSON contenente la ground truth.

La *ground truth* così costruita copre l'insieme dei termini clinici contenuti in ciascun *chunk*. Per ogni termine sono stati considerati validi più concetti candidati, in modo da garantire che il *concept extractor* ne rilevi correttamente almeno uno.

Ad esempio, per il termine estratto "*breathlessness*", sono stati considerati validi più concetti SNOMED CT corrispondenti, come mostra la Figura 41.

```
"SNOMEDCTConcepts": [
            {
                "Description": "Dyspnea (finding)",
                "Code": "267036007",
                "Score": 0.16991747915744781
            },
            {
                "Description": "Increasing breathlessness (finding)",
                "Code": "297216006",
                "Score": 0.06654880940914154
            },
            {
                "Description": "Difficulty breathing (finding)",
                "Code": "230145002",
                "Score": 0.062349770218133926
            },
            {
                "Description": "Dyspnea on exertion (finding)",
                "Code": "60845006",
                "Score": 0.05510967597365379
            }
        ]
```


Figura 41: concetti validati per il termine individuato "Breathlessness"

Per valutare in che misura sia stato necessario il processo di validazione manuale e quanto lo strumento automatico abbia contribuito alla costruzione della *ground truth*, è stato analizzato il comportamento di Comprehend Medical durante la validazione. I dati, riportati per ciascun chunk in Tabella 12 (Appendice 2), riguardano i concetti estratti da *Comprehend Medical*, quelli eliminati perché errati, quelli aggiunti a mano e i concetti validati che compongono la *ground truth* finale. Poiché estrazione e validazione sono state condotte separatamente per ciascun *chunk*, l'analisi è stata svolta a livello del singolo *chunk*. I risultati relativi all'intera linea guida sono stati quindi ottenuti come media dei risultati individuali.

In primo luogo, si è valutata l'accuratezza di *Comprehend Medical*: per ciascun *chunk* si è calcolata la percentuale di concetti errati restituiti dallo strumento, cioè eliminati in fase di validazione, rispetto al totale dei concetti che lo strumento aveva estratto. La distribuzione di questi valori, raffigurata dal box plot in Figura 42, mostra un valore mediano di 0,71 (linea centrale del box) e una media di 0,66 (indicata dalla X), con un intervallo interquartile compreso tra 0,55 e 0,78. Sono presenti due valori anomali isolati di 0,14 e 0,17, corrispondenti a *chunk* con

un’estrazione di *Comprehend Medical* particolarmente accurata rispetto alla media. La distribuzione, relativamente compatta e con la mediana superiore alla media, suggerisce un'asimmetria verso il basso: la maggior parte dei *chunk* presenta un tasso di errore elevato. (Figura 42)

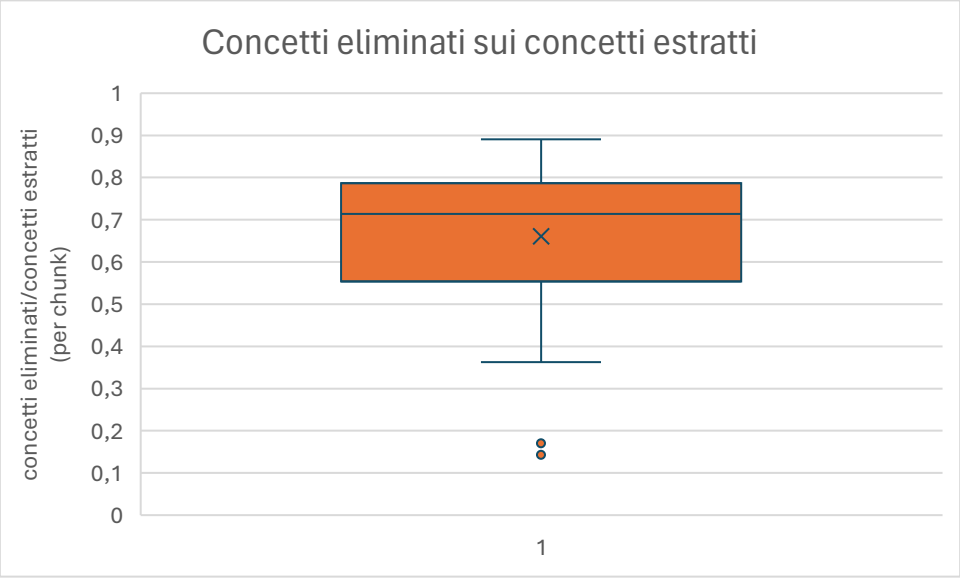


Figura 42: box plot della distribuzione, sui chunk della linea guida, della frazione di concetti eliminati in fase di validazione rispetto ai concetti estratti da Comprehend Medical. La linea al centro del riquadro indica la mediana, la croce la media, i bordi del riquadro il primo e il terzo quartile; i baffi arrivano ai valori minimo e massimo non anomali, i punti isolati sono valori anomali.

Aggregando questi dati sull’intero corpus (Figura 43), risulta che il 66% dei concetti estratti da *Comprehend Medical* è stato eliminato in quanto errato, a fronte di un 34% correttamente mantenuto. Questo fenomeno è riconducibile alla caratteristica di *Comprehend Medical* di restituire i cinque concetti più probabili per ciascun termine clinico individuato[100]. Tale modalità ha inevitabilmente prodotto un numero elevato di concetti solo vagamente riconducibili al termine originario e, dunque, errati. Questa caratteristica ha aumentato la probabilità che, tra i cinque concetti candidati, almeno uno risulti corretto. Risulta, tuttavia, imprescindibile una validazione umana successiva.

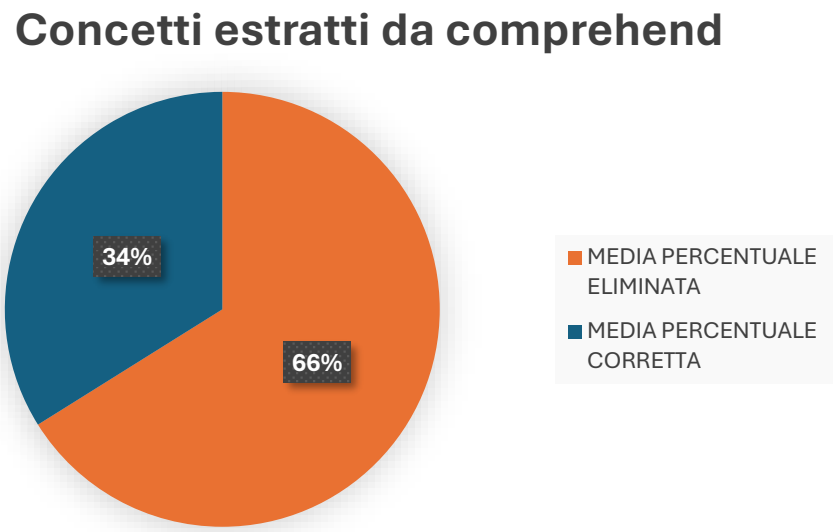


Figura 43: grafico a torta della media percentuale dei concetti eliminati in fase di validazione.

In secondo luogo, si è valutata la completezza di Comprehend Medical: per ciascun chunk si è calcolata la percentuale di concetti aggiunti manualmente rispetto al totale dei concetti validati che compongono la *ground truth* finale.

L'aggiunta manuale è parte del metodo, e non una sua correzione. La *ground truth* è il riferimento con cui si valuta l'estrattore, quindi, non deve dipendere dai limiti dello strumento usato per costruirla.

La distribuzione, raffigurata nel box plot di Figura 44, è marcatamente più concentrata su valori bassi, con una mediana di circa 0,07 e una media di circa 0,13, segno che la maggioranza dei *chunk* ha richiesto un'integrazione manuale minima. Si osservano tre valori anomali più elevati, corrispondenti a *chunk* in cui l'estrazione automatica è risultata particolarmente carente, rendendo necessaria un'integrazione manuale sostanziale.

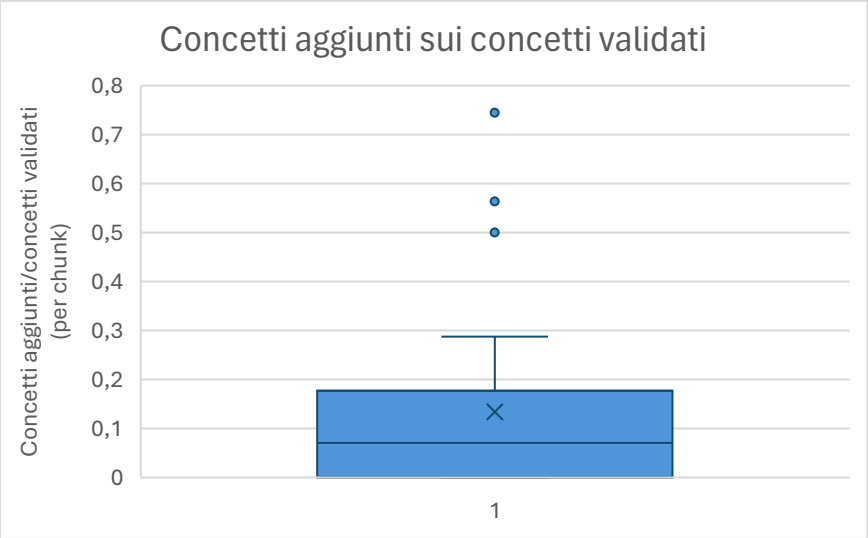


Figura 44: box plot della distribuzione, sui chunk della linea guida, della frazione di concetti aggiunti in fase di validazione rispetto ai validati della ground truth. La linea al centro del riquadro indica la mediana, la croce la media, i bordi del riquadro il primo e il terzo quartile; i baffi arrivano ai valori minimo e massimo non anomali, i punti isolati sono valori anomali.

Aggregando questi dati sull'intero corpus (Figura 45), la composizione finale della ground truth risulta costituita per il 13% da concetti aggiunti manualmente e per l'87% da concetti correttamente estratti da *Comprehend Medical*. Questo metodo ibrido, automatico più validazione manuale, ha permesso di costruire una *ground truth* affidabile in tempi ridotti, supplendo all'assenza di una supervisione clinica specialistica nel presente lavoro.

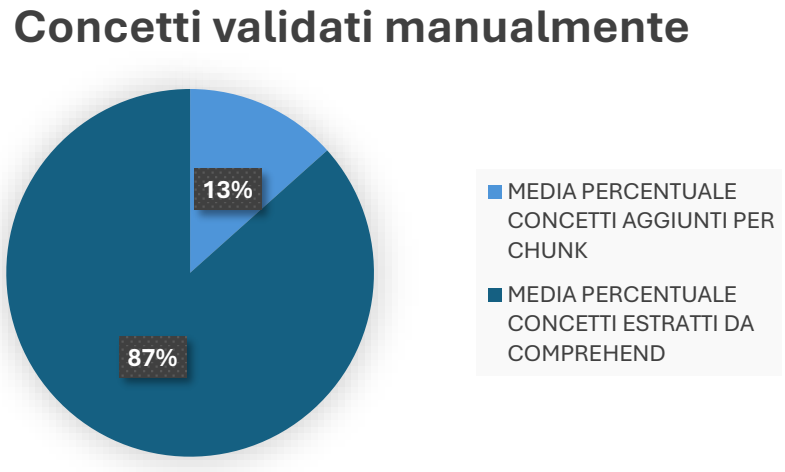


Figura 45: grafico a torta della media percentuale dei concetti aggiunti ed estratti utilizzando Comprehend Medical sul totale dei concetti validati manualmente.

5.6.3 Performance dell'estrattore di concetti

I risultati sperimentali, ottenuti confrontando le due strategie di riordinamento nell'estrattore di concetti, un modello *cross-encoder* (neuml/biomedbert-base-reranker) e un LLM (Ministral, modello mistral.ministral-3-3b-instruct), su entrambi i dizionari costruiti e trattati nel paragrafo 5.6.1 (dizionario completo SNOMED_dict e quello filtrato tramite *Knowledge Graph* COPD_dict), sono riportati in Tabella 10. Le metriche sono state calcolate confrontando i concetti estratti dall'estrattore e la *ground truth* discussa al paragrafo 5.6.2.

modelli	Cross encoder		LLM	
Dizionari	SNOMED_dict	COPD_dict	SNOMED_dict	COPD_dict
Precision	0,33	0,46	0,58	0,72
Recall	0,12	0,11	0,59	0,46
F1-score	0,17	0,17	0,58	0,56

Tabella 10: Risultati sperimentali al variare delle strategie di riordinamento (cross-encoder ed LLM) e dei dizionari (SNOMED_dict e COPD_dict).

Coerentemente con l'approccio metodologico adottato per la creazione e la valutazione della *ground truth*, *precision*, *recall* e *F1-score* sono state calcolate separatamente per ciascun *chunk*, e i valori riportati in tabella rappresentano la media di tali metriche sui singoli *chunk*.

Un primo dato che emerge con chiarezza è la netta superiorità dell'LLM (Ministral 3B) rispetto al *cross-encoder* (BiomedBERT reranker) su tutte e tre le metriche, indipendentemente dal dizionario impiegato. Ciò suggerisce che l'LLM presenta una capacità di comprensione contestuale più ampia che sembra favorire un'identificazione più accurata del candidato corretto anche in presenza di formulazioni linguistiche ambigue o sinonimiche, coerentemente con quanto discusso a livello teorico nel paragrafo 4.4.

Per quanto riguarda l'effetto del dizionario, si osserva un comportamento diverso a seconda del modello di riordinamento impiegato. Con il *cross-encoder*, il

passaggio dal dizionario completo a quello filtrato per la BPCO migliora la *precision*, poiché la riduzione dello spazio dei candidati elimina concetti chiaramente estranei al dominio, riducendo i falsi positivi; la *recall*, tuttavia, rimane sostanzialmente invariata e la *F1-score* risulta identica in entrambe le configurazioni, segno che il collo di bottiglia delle prestazioni non risiede nella dimensione del dizionario, bensì nella capacità discriminativa del modello di riordinamento stesso.

Con l'LLM, invece, il restringimento del dizionario produce un effetto più marcato e in parte opposto: la *precision* aumenta sensibilmente, ma la *recall* cala, con una *F1-score* leggermente inferiore nella configurazione con COPD_dict. Questo risultato è ragionevole se si considera che il dizionario COPD_dict, filtrato entro tre gradi di separazione dal concetto COPD nel *Knowledge Graph*, può escludere concetti clinicamente rilevanti ma topologicamente più distanti (ad esempio comorbidità o farmaci non direttamente connessi al concetto di BPCO), penalizzando così la *recall*; al contempo, la riduzione del rumore nello spazio dei candidati permette all'LLM di essere più selettivo e preciso nelle scelte residue.

In sintesi, la configurazione che massimizza la F1-score complessiva è quella LLM + SNOMED_dict (0,58), che rappresenta il miglior compromesso tra *precision* e *recall*. Il cross-encoder, pur presentando un vantaggio in termini di costo computazionale ed efficienza, mostra in questo caso prestazioni nettamente inferiori in entrambe le configurazioni.

Va infine osservato che, pur rappresentando il miglior risultato ottenuto, una *F1-score* compresa tra 0,56 e 0,58 lascia ancora margine di miglioramento significativo, coerentemente con la complessità terminologica e l'elevata ambiguità semantica del dominio clinico discusse nei capitoli precedenti, e con le difficoltà intrinseche di costruzione di una *ground truth* completa ed esaustiva, evidenziate nel paragrafo 5.6.2.

Si riporta di seguito un esempio concreto di estrazione su una porzione di chunk, ottenuto utilizzando la configurazione che massimizza i risultati (LLM+SNOMED_dict).

Nel testo sono evidenziati i termini clinici: in verde i termini estratti correttamente e in giallo i termini presenti nella ground truth ma non estratti.

> Initial pharmacotherapy should be based on the patient's GOLD group (Figure 3.8). Patients should be offered guidance and follow-up on self-management of breathlessness, and stress management, and be given a written action plan. Comorbidities should be managed as per specific guidelines, irrespective of the presence of COPD (Chapter 5). Follow-up pharmacotherapy should be based on the patient's symptoms (dyspnea or exacerbations) at review.

La Tabella 11 confronta, termine per termine, i concetti estratti con quelli considerati validi nella *ground truth*.

Termine nel testo	Concetti nella ground truth	Concetto estratto	Esito
Pharmacotherapy	Drug therapy status (410684002); **Supportive pharmacotherapy (169414008)**; Drug therapy observable (363819003)	169414008	Corretto
Follow-up	**Follow-up (308273005)**	308273005	Corretto
Self-management	Respiratory health self-management education (427625003); Asthma self-management (699728000)	management (393450003)	Non riconosciuto; estratto un termine generico
breathlessness	**Dyspnea (267036007)**; Increasing breathlessness (297216006); Dyspnea on exertion (60845006); Difficulty breathing (230145002)	267036007	Corretto
stress management	**Stress management (procedure) (226060000)** e altri 3 concetti	226060000	Corretto
COPD	**Chronic obstructive lung disease (13645005)** e altri 5 concetti	13645005	Corretto
comorbidities	Co-morbid conditions (398192003)	-	Non riconosciuto
Symptoms	Multiple symptoms (38276004) e altri 3 concetti	-	Non riconosciuto
exacerbations	**Acute exacerbation of chronic obstructive airways disease (195951007)** e altri 2 concetti	195951007	Corretto
Dyspnea	**Dyspnea (267036007)**; Increasing breathlessness (297216006); Dyspnea on exertion (60845006); Difficulty breathing (230145002)	267036007	Corretto

Tabella 11: confronto tra concetti estratti e ground truth sulla porzione di chunk in analisi. In grassetto i concetti della ground truth individuati dall'estrattore.

Su 10 occorrenze di termini clinici, 7 sono estratte correttamente. Gli errori sono dovuti al non riconoscimento dei termini*: comorbidities* e *symptoms*, a cui si aggiunge *self-management*, di cui l'estrattore coglie solo la componente generica

management. Quest'ultima perdita è rilevante, perché il concetto di autogestione è clinicamente diverso dalla gestione in generale.

Va infine osservato che la *ground truth* (28 concetti riportati in Tabella 13, Appendice 3) è molto più ampia dei concetti estratti (7 riportati all'Appendice 3), coerentemente con la scelta del paragrafo 5.6.2 di considerare validi più concetti per ciascun termine, così da incrementare la probabilità di validare i concetti estratti.

6 Conclusioni e sviluppi futuri

6.1 Conclusioni

Il progetto GUIDE-AI si è posto l'obiettivo di studiare, processare e formalizzare la conoscenza clinica contenuta nelle linee guida per la gestione farmacologica delle malattie croniche, costruendo un sistema di supporto alla decisione clinica ancorato a una base di conoscenza costituita da fonti certificate e redatte da esperti del dominio, anziché a un modello puramente generativo e probabilistico.

Se l'obiettivo generale di GUIDE-AI è aumentare l'aderenza alle linee guida nella gestione delle malattie croniche a livello europeo, il contributo specifico di questa tesi si è concentrato sulla BPCO, adottando un approccio basato sulla conoscenza al fine di evitare i rischi di allucinazione, obsolescenza e scarsa trasparenza tipici dei modelli generativi in un contesto ad alto rischio come quello sanitario.

I capitoli della tesi hanno approfondito questi temi procedendo dal generale al particolare. Sono stati analizzati i due paradigmi di intelligenza artificiale: i sistemi esperti e l'architettura *Retrieval-Augmented Generation* (RAG), nelle sue componenti di rappresentazione della conoscenza e di generazione tramite LLM. Si è passati per l'approfondimento delle modalità di rappresentazione della conoscenza clinica, a partire dal concetto di ontologia e dalle terminologie standardizzate SNOMED CT e ATC, fino alle due strutture dati alla base dei sistemi RAG moderni: il *Knowledge Graph* e il database vettoriale. Per finire il quadro teorico, è stata presentata l'architettura dell'estrattore di concetti, applicato al fine di associare ai nodi del grafo di conoscenza relativo alla BPCO i corrispondenti termini SNOMED CT. La struttura si articola in un percorso basato su dizionario, un percorso neurale basato su GLiNER2 e recupero denso, e una fase di disambiguazione tramite riordinamento, confrontando le due strategie alternative di *cross-encoder* e LLM. È stato inoltre definito il quadro metodologico di valutazione, tramite la costruzione di una *ground truth* di riferimento e l'adozione di metriche di valutazione.

Il capitolo 5 è stato dedicato alla presentazione dell'implementazione pratica del sistema e dei risultati sperimentali ottenuti. È stato descritto il processo di formalizzazione del percorso terapeutico della BPCO; tale formalizzazione è stata successivamente tradotta in un sistema esperto eseguibile tramite il *Composer* PROforma. Nonostante gli evidenti punti di forza dell'approccio tradizionale, come la trasparenza, la spiegabilità dell'algoritmo decisionale e l'assenza di allucinazioni, la formalizzazione con il linguaggio PROforma presenta alcuni limiti, come l'onerosità dell'implementazione e la rigidità della struttura che non permette di gestire scenari clinici complessi. Per superare tali limiti si è avviato lo sviluppo di un'architettura RAG, di cui questo lavoro realizza le basi di conoscenza. È stata, quindi, descritta la costruzione del *Knowledge Graph*, a partire dalla selezione delle sezioni rilevanti della linea guida GOLD fino alla sua integrazione con SNOMED CT. Infine, è stata presentata l'implementazione concreta del *Concept Extractor*, comprendente la costruzione dei dizionari terminologici, la creazione della *ground truth* tramite un approccio ibrido automatico-manuale basato su *AWS Comprehend Medical*, e la valutazione sperimentale delle due strategie di riordinamento: i risultati hanno mostrato una netta superiorità dell'LLM (Ministral 3B) rispetto al *cross-encoder* (BiomedBERT), con una *F1-score* massima di 0,58 nella configurazione LLM + SNOMED_dict.

Il lavoro di tesi presenta numerosi punti di forza. Il primo è la formalizzazione completa e verificabile del percorso terapeutico della BPCO, dalla presa in carico al follow-up. La definizione delle variabili cliniche, condotta con il confronto di esperti clinici, ha prodotto un sistema esperto eseguibile e un modulo strutturato di raccolta dati per i medici degli ospedali partner, che costituisce il gold standard per valutare gli approcci di estrazione automatica nel progetto. Il secondo è il *Knowledge Graph*, che conserva la struttura originaria della linea guida (sezioni, chunk, figure) e si propone di collegarla ai concetti SNOMED CT, a loro volta legati a concetti ATC e ad una classificazione per classi farmacologiche creata ad

hoc: ogni informazione risulterà così tracciabile fino alla fonte, in linea con il principio di ancorare il sistema ad una base di conoscenza. La conversione delle figure in testo strutturato, sottoposta a doppio controllo, ha reso utilizzabile dal RAG anche la conoscenza contenuta nei flowchart e nelle tabelle. Il terzo è l'estrattore di concetti, già sviluppato per un altro dominio (l'artrite reumatoide) e adattato alla BPCO con nuovi dizionari e una nuova *ground truth*, il che suggerisce una buona riutilizzabilità. L'applicazione di un estrattore di concetti dedicato vincola l'estrazione ad una terminologia di riferimento.

Per contro, si rilevano numerose limitazioni. Il sistema esperto presenta i limiti tipici di questo paradigma. La formalizzazione è manuale e onerosa, richiede esperti e va ripetuta a ogni aggiornamento della linea guida, con costi di manutenzione che crescono con il numero di patologie. La struttura è rigida: il sistema gestisce solo i percorsi previsti, parte dall'ipotesi di una diagnosi di BPCO già confermata e non modella la multimorbidità, che è invece uno degli obiettivi di GUIDE-AI. Richiede inoltre dati strutturati in ingresso e non legge direttamente i referti in testo libero.

Sia la conversione delle figure in XML che la *ground truth* sono state validate dall'autrice, non da clinici, quindi, eventuali errori di interpretazione clinica potrebbero non essere stati intercettati. Questo lavoro costruisce inoltre le basi di conoscenza del RAG, ma non ne implementa né valuta il funzionamento end-to-end: i vantaggi attesi rispetto al sistema esperto restano di natura teorica e non sono stati misurati.

Nel complesso, il lavoro svolto ha evidenziato l'importanza e la fattibilità di un approccio basato sulla conoscenza in ambito clinico e ne chiarisce i compromessi. Il sistema esperto offre trasparenza e spiegabilità, al prezzo di rigidità e di manutenzione costante. Il RAG, di cui sono state costruite le basi, promette maggiore flessibilità, ma richiede una valutazione sperimentale completa. I limiti descritti non annullano i risultati ottenuti, ma indicano le priorità per le fasi successive di GUIDE-AI.

6.2 Sviluppi futuri

Il presente lavoro di tesi ha affrontato una porzione specifica dell'architettura complessiva del progetto GUIDE-AI, in particolare le fasi di formalizzazione della conoscenza (sistema esperto tramite *Composer*, *Knowledge Graph*) e di estrazione automatica dei concetti clinici da linea guida (*Concept Extractor*). Diversi componenti dell'architettura complessiva, pur presentati a livello teorico nei capitoli precedenti, non sono stati oggetto di implementazione diretta in questo lavoro, e costituiscono pertanto naturali sviluppi futuri del progetto.

Il primo sviluppo del lavoro prevede il collegamento, all'interno del grafo di conoscenza, dei concetti SNOMED CT ai *chunk* della linea guida. Una volta identificate, attraverso l'estrattore di concetti, le entità cliniche presenti nell'attributo "text" di ciascun *chunk*, è possibile mappare questi ultimi tramite la relazione HAS_CONCEPT ai corrispondenti concetti SNOMED CT. In questo modo, ogni porzione di testo delle linee guida non solo mantiene la propria struttura informativa, ma risulta anche ancorata a concetti clinici normalizzati, garantendo maggiore coerenza semantica e interoperabilità.

In secondo luogo, resta da implementare e valutare l'intera pipeline RAG nella sua componente di *query processing*, recupero di *chunk* rilevanti dal *Knowledge Graph* e dal database vettoriale, e generazione finale della risposta.

In terzo luogo, l'architettura originaria del progetto GUIDE-AI prevede un modulo dedicato alla gestione della multimorbidità pensato per gestire le interazioni tra piani terapeutici multipli nel paziente con più patologie croniche concomitanti. Questo modulo non è stato affrontato in questo lavoro e rimane un'estensione per gli sviluppi futuri.

In quarto luogo, il presente lavoro si è concentrato esclusivamente sulla BPCO, sia nella formalizzazione del sistema esperto sia nella costruzione del grafo di conoscenza e dell'estrattore di concetti. Sarà quindi necessario estendere la

metodologia sviluppata alle altre patologie croniche previste dal progetto GUIDE-AI: scompenso cardiaco cronico (SCC), malattia renale cronica (MRC) e asma.

In quinto luogo, dal punto di vista specifico dell'estrattore di concetti, i risultati ottenuti (*F1-score* massima di 0,58) lasciano margine di miglioramento significativo: sviluppi futuri potrebbero riguardare la sperimentazione di modelli di riordinamento al fine di colmare il divario ancora presente tra le prestazioni ottenute e quelle auspicabili per un utilizzo clinico affidabile.

Un ulteriore sviluppo riguarda la validazione. È consigliato coinvolgere clinici nella revisione della *ground truth* e valutare il sistema esperto su un numero maggiore di casi reali. Servirebbe inoltre un confronto sperimentale tra sistema esperto e RAG sugli stessi casi, in termini di correttezza, spiegabilità e onere di manutenzione.

Appendice 1 – Prompt per traduzione delle figure in formato XML

```
Ti allego una figura tratta dal report GOLD 2026 sulla BPCO.
Trascrivila in formato XML, in modo che il suo contenuto possa
essere letto e indicizzato da un sistema automatico.

Regole:
- Includi tutto il testo presente nella figura, senza
riassumerlo né tradurlo, e senza aggiungere informazioni che non
compaiono.
- Rappresenta la struttura. Per un flowchart: i nodi, le
condizioni sulle diramazioni e le azioni finali, collegati
tramite identificativi (id) e transizioni. Per una tabella:righe
e colonne con le relative intestazioni.
- Riporta le note a piè di pagina e i rimandi ad altre figure.
- Usa nomi di elementi e di attributi in inglese, chiari e
coerenti.
- Restituisci solo il documento XML, ben formato, senza commenti
o spiegazioni.
```

Appendice 2 – Dati riguardanti la creazione della ground truth

Titolo del chunk	# concetti SNOMED CT unici estratti da comprehend	# concetti validati	# concetti aggiunti	# concetti eliminati
PHARMACOLOGICAL MAINTENANCE TREATMENT OF COPD	231	112	0	119
Initial pharmacological management	169	49	0	120
Follow-up pharmacological management	241	115	12	138
PHARMACOLOGICAL TREATMENT_ Bronchodilators	149	73	12	88
PHARMACOLOGICAL TREATMENT_ Glucocorticoids	102	68	3	37
PHARMACOLOGICAL TREATMENT_ Antibiotics	268	165	28	131
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Maintenance therapy for stable COPD	116	56	0	60
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Bronchodilators	476	200	45	321
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Methylxanthines	197	102	14	109

Titolo del chunk	# concetti SNOMED CT unici estratti da comprehend	# concetti validati	# concetti aggiunti	# concetti eliminati
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Combination bronchodilator therapy	205	63	5	147
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Anti-inflammatory agents	20	10	5	15
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Inhaled corticosteroids	388	133	75	330
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Triple therapy (LABA+LAMA+ICS)	127	25	0	102
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Oral glucocorticoids	101	45	0	56
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Phosphodiesterase-4 inhibitors	88	75	2	15
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Antibiotics	75	27	4	52

Titolo del chunk	# concetti SNOMED CT unici estratti da comprehend	# concetti validati	# concetti aggiunti	# concetti eliminati
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Mucolytic and antioxidant agent	75	25	3	53
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Phosphodiesterase-3 and -4 inhibitors	108	35	1	74
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Biologic therapy	125	41	6	90
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Other drugs with potential to reduce exacerbations_	152	50	10	112
APPENDIX 3 – OVERVIEW OF THE EVIDENCE_ PHARMACOTHERAPY _ Adherence to inhaled COPD medications	146	40	2	108
Figure 3.7	35	30	0	5
Figure 3.8	164	27	0	137
Figure 3.9	178	42	0	136
Figure 3.10	35	10	2	27
Figure 3.11	193	32	2	163
Figure 3.12	73	8	0	65
Figure A3.1	88	47	35	76
Figure A3.2	178	40	1	139
Figure A3.3	299	73	21	247

Tabella 12: Quantizzazione del processo di costruzione e validazione della ground truth.

Appendice 3 – Concetti estratti e validati d’esempio

Concetti estratti dell’esempio al 5.6.3


```
{'start': 796, 'end': 811, 'concept_term': ' Supportive
pharmacotherapy', 'semantic_tag': 'regime/therapy', 'term_type':
'SYN', 'concept_id': '169414008'}

{'start': 569, 'end': 578, 'concept_term': 'follow-up',
'semantic_tag': "['wait and see', 'wait and see', 'finding']",
'term_type': 'FSN', 'concept_id': '308273005'}

{'start': 601, 'end': 615, 'concept_term': ' Breathlessness',
'semantic_tag': 'finding', 'term_type': 'SYN', 'concept_id':
'267036007'}

{'start': 621, 'end': 638, 'concept_term': 'stress management',
'semantic_tag': "['procedure']", 'term_type': 'FSN',
'concept_id': '226060000'}

{'start': 855, 'end': 862, 'concept_term': 'dyspnea',
'semantic_tag': "['finding']", 'term_type': 'FSN', 'concept_id':
'267036007'}

{'start': 425, 'end': 435, 'concept_term': 'management',
'semantic_tag': "['procedure']", 'term_type': 'FSN',
'concept_id': '393450003'}

{'start': 768, 'end': 772, 'concept_term': ' exacerbation of
your COPD', 'semantic_tag': '', 'term_type': 'ANNOTATION',
'concept_id': '195951007'}

{'start': 1826, 'end': 1830, 'concept_term': ' CAO - Chronic
airflow obstruction', 'semantic_tag': 'disorder', 'term_type':
'SYN', 'concept_id': '13645005'}
```

Ground truth dell'esempio al paragrafo 5.6.3

Concetti	Codici
'Drug therapy status (situation)','Supportive pharmacotherapy (regime/therapy)','Drug therapy observable (observable entity)', 'Follow-up', 'Respiratory health self-management education (procedure)','Asthma self-management (regime/therapy)','Dyspnea (finding)','Increasing breathlessness (finding)','Dyspnea on exertion (finding)','Difficulty breathing (finding)', 'Stress management (procedure)','Stress management (qualifier value)','Stress management surveillance (regime/therapy)', 'Stress management education, guidance, and counseling (procedure)','Chronic obstructive lung disease (disorder)','Chronic obstructive bronchitis (disorder)','Chronic obstructive pulmonary disease screening (procedure)','Chronic lung disease (disorder)','Chronic obstructive pulmonary disease clinic (environment)','Chronic obstructive pulmonary disease monitoring (regime/therapy)', 'Co-morbid conditions (finding)','Multiple symptoms (finding)','General symptom (finding)','Symptom severity (finding)','Associated symptom (finding)', 'Acute exacerbation of chronic obstructive airways disease (disorder)','Acute exacerbation of chronic asthmatic bronchitis (disorder)','Exacerbation of asthma (disorder)'	410684002, 169414008, 363819003, 308273005, 427625003, 699728000, 267036007, 297216006, 60845006, 230145002, 226060000, 441151000124104, 410418004, 410315002, 13645005, 185086009, 171230000, 413839001, 702839006, 390891009, 398192003, 38276004, 267022002, 162465004, 102483000, 195951007, 442025000, 281239006

Tabella 13: ground truth della porzione di chunk in analisi.

Bibliografia

[1] «GUIDE-AI | Iniziativa IHI per la salute innovativa». Consultato: 12 ottobre 2026. [Online]. Disponibile su: http://www.ihi.europa.eu/projects-results/project-factsheets/guide-ai

[2] «GUIDE-AI – Guiding your treatment with current evidence | GUIDE-AI | Project | Fact Sheet | HORIZON», CORDIS | European Commission. Consultato: 7 settembre 2026. [Online]. Disponibile su: https://cordis.europa.eu/project/id/101253015

[3] «Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease», Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2026, 2026.

[4] B. Celli *et al.*, «Definition and Nomenclature of Chronic Obstructive Pulmonary Disease: Time for Its Revision», *Am. J. Respir. Crit. Care Med.*, vol. 206, fasc. 11, pp. 1317–1325, dic. 2022, doi: 10.1164/rccm.202204-0671PP.

[5] I. Yang, C. Jenkins, e S. Salvi, «Chronic obstructive pulmonary disease in never-smokers: risk factors, pathogenesis, and implications for prevention and treatment», Lancet Respir Med, vol. 10, fasc. 5, pp. 497–511, 2022.

[6] P. Chen, Y. Li, D. Wu, F. Liu, e C. Cao, «Secondhand Smoke Exposure and the Risk of Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis», *Int. J. Chron. Obstruct. Pulmon. Dis.*, vol. Volume 18, pp. 1067–1076, giu. 2023, doi: 10.2147/COPD.S403158.

[7] M. Orozco-Levi, J. Garcia-Aymerich, J. Villar, A. Ramírez-Sarmiento, J. M. Antó, e J. Gea, «Wood smoke exposure and risk of chronic obstructive pulmonary disease», *Eur. Respir. J.*, vol. 27, fasc. 3, pp. 542–546, mar. 2006, doi: 10.1183/09031936.06.00052705.

[8] L. M. Paulin *et al.*, «Occupational Exposures Are Associated with Worse Morbidity in Patients with Chronic Obstructive Pulmonary Disease», *Am. J. Respir. Crit. Care Med.*, vol. 191, fasc. 5, pp. 557–565, mar. 2015, doi: 10.1164/rccm.201408-1407OC.

[9] «GBD Compare», Institute for Health Metrics and Evaluation. Consultato: 19 settembre 2026. [Online]. Disponibile su: https://vizhub.healthdata.org/gbd-compare

[10] M. H. Cho, B. D. Hobbs, e E. K. Silverman, «Genetics of chronic obstructive pulmonary disease: understanding the pathobiology and heterogeneity of a complex disorder», *Lancet Respir. Med.*, vol. 10, fasc. 5, pp. 485–496, mag. 2022, doi: 10.1016/S2213-2600(21)00510-5.

[11] J. Stoller e L. Aboussouan, «Alpha1-antitrypsin deficiency», Lancet, vol. 365, fasc. 9478, 2005.

[12] American Thoracic Society (ATS), «Surveillance for respiratory hazards in the occupational setting», vol. 126, pp. 952–956, 1982.

[13] P. W. Jones, G. Harding, P. Berry, I. Wiklund, W.-H. Chen, e N. Kline Leidy, «Development and first validation of the COPD Assessment Test», *Eur. Respir. J.*, vol. 34, fasc. 3, pp. 648–654, set. 2009, doi: 10.1183/09031936.00102509.

[14] P. W. Jones *et al.*, «Validity of the Chronic Airways Assessment Test (CAAT) in asthma, asthma+COPD and COPD in NOVELTY», *ERJ Open Res.*, vol. 11, fasc. 4, pp. 01359–02024, lug. 2025, doi: 10.1183/23120541.01359-2024.

[15] R. Buhl, M. Miravitlles, A. Anzueto, e S. Brunton, «Long-acting muscarinic antagonist and long-acting β_2 -agonist combination for the treatment of maintenance therapy–naïve patients with chronic obstructive pulmonary disease: a narrative review», *Ther. Adv. Respir. Dis.*, vol. 18, p. 17534666241279115, gen. 2024, doi: 10.1177/17534666241279115.

[16] D. A. Lipson *et al.*, «Once-Daily Single-Inhaler Triple versus Dual Therapy in Patients with COPD», *N. Engl. J. Med.*, vol. 378, fasc. 18, pp. 1671–1680, mag. 2018, doi: 10.1056/NEJMoa1713901.

[17] K. F. Rabe *et al.*, «Triple Inhaled Therapy at Two Glucocorticoid Doses in Moderate-to-Very-Severe COPD», *N. Engl. J. Med.*, vol. 383, fasc. 1, pp. 35–48, lug. 2020, doi: 10.1056/NEJMoa1916046.

[18] D. Singh *et al.*, «Blood Eosinophils and Chronic Obstructive Pulmonary Disease: A Global Initiative for Chronic Obstructive Lung Disease Science Committee 2022 Review», *Am. J. Respir. Crit. Care Med.*, vol. 206, fasc. 1, pp. 17–24, lug. 2022, doi: 10.1164/rccm.202201-0209PP.

[19] IOM (Institute of Medicine), *Clinical Practice Guidelines We Can Trust*. Washington, DC: The National Academies Press, 2011.

[20] *Legge* 8 marzo 2017, n.24

[21] B. G. Buchanan e R. G. Smith, «Fundamentals of Expert Systems», *Annu. Rev. Comput. Sci.*, vol. 3, fasc. 1, pp. 23–58, giu. 1988, doi: 10.1146/annurev.cs.03.060188.000323.

[22] J. Fox, N. Johns, C. Lyons, A. Rahmanzadeh, R. Thomson, e P. Wilson, «PROforma: a general technology for clinical decision support systems», *Comput. Methods Programs Biomed.*, vol. 54, fasc. 1–2, pp. 59–67, set. 1997, doi: 10.1016/S0169-2607(97)00034-5.

[23] «Informazioni su Deontics». Consultato: 12 ottobre 2026. [Online]. Disponibile su: https://deontics.com/about/

[24] Bury Jonathan, Fox John, e Sutton David, «The PROforma Guideline Specification Language: Progress and Prospects», in *Studies in Health Technology and Informatics*, IOS Press, 2001. doi: 10.3233/978-1-60750-927-1-13.

[25] D. R. Sutton e J. Fox, «The Syntax and Semantics of the PRO *forma* Guideline Modeling Language», *J. Am. Med. Inform. Assoc.*, vol. 10, fasc. 5, pp. 433–443, set. 2003, doi: 10.1197/jamia.M1264.

[26] J. Fox, V. Patkar, e R. Thomson, «Decision support for health care: the PROforma evidence base», *J. Innov. Health Inform.*, vol. 14, fasc. 1, pp. 49–54, mar. 2006, doi: 10.14236/jhi.v14i1.616.

[27] D. Jurafsky e J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*, Second. Prentice Hall, 2009.

[28] R. Grishman e B. Sundheim, «DESIGN OF THE MUC-6 EVALUATION».

[29] W. Shen, J. Wang, e J. Han, «Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions», *IEEE Trans. Knowl. Data Eng.*, vol. 27, fasc. 2, pp. 443–460, feb. 2015, doi: 10.1109/TKDE.2014.2327028.

[30] P. Rajpurkar, J. Zhang, K. Lopyrev, e P. Liang, «SQuAD: 100,000+ Questions for Machine Comprehension of Text», in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

[31] W. J. Hutchins, *Machine translation: past, present, future*. in Ellis Horwood series in computers and their applications. Chichester [West Sussex] : New York: Ellis Horwood ; Halsted Press, 1986.

[32] J. Weizenbaum, «ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine», *Commun. ACM*, vol. 9, fasc. 1, pp. 36–45, gen. 1966, doi: 10.1145/365153.365168.

[33] C. D. Manning, «Foundations of Statistical Natural Language Processing».

[34] Y. Bengio, R. Ducharme, P. Vincent, e C. Jauvin, «A Neural Probabilistic Language Model».

[35] A. Vaswani *et al.*, «Attention is All you Need», NeurIPS 2017

[36] H. Naveed *et al.*, «A Comprehensive Overview of Large Language Models», *ACM Trans. Intell. Syst. Technol.*, vol. 16, fasc. 5, pp. 1–72, ott. 2025, doi: 10.1145/3744746.

[37] R. J. Boxer, «Advantages and utilization of telemedicine», *mHealth*, vol. 5, pp. 12–12, mag. 2019, doi: 10.21037/mhealth.2019.04.02.

[38] S. J. Russell e P. Norvig, *Artificial intelligence: a modern approach*, 13ª ed. in Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall, 1995.

[39] J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», NAACL 2019

[40] Y. Liu *et al.*, «RoBERTa: A Robustly Optimized BERT Pretraining Approach», 26 luglio 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[41] OpenAI *et al.*, «GPT-4 Technical Report», 4 marzo 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[42] G. Team *et al.*, «Gemini: A Family of Highly Capable Multimodal Models», 9 maggio 2025, *arXiv*: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805.

[43] C. Raffel *et al.*, «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer», JMLR

[44] M. Lewis *et al.*, «BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.

[45] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, e D. S. W. Ting, «Large language models in medicine», *Nat. Med.*, vol. 29, fasc. 8, pp. 1930–1940, ago. 2023, doi: 10.1038/s41591-023-02448-8.

[46] Regolamento (UE) 2016/679, *GDPR*.

[47] L. Cheng, F. Liu, e D. (Daphne) Yao, «Enterprise data breach: causes, challenges, prevention, and future directions», *WIREs Data Min. Knowl. Discov.*, vol. 7, fasc. 5, p. e1211, set. 2017, doi: 10.1002/widm.1211.

[48] E. Derner, K. Batistič, J. Zahálka, e R. Babuška, «A Security Risk Taxonomy for Prompt-Based Interaction With Large Language Models», *IEEE Access*, vol. 12, pp. 126176–126187, 2024, doi: 10.1109/ACCESS.2024.3450388.

[49] Regolamento (UE) 2024/1689, *AI Act*.

[50] P. Lewis *et al.*, «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks», NeurIPS 2020.

[51] S. Das, S. Saha, e R. Srihari, «Diving Deep into Modes of Fact Hallucinations in Dialogue Systems», in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 684–699. doi: 10.18653/v1/2022.findings-emnlp.48.

[52] C. Wang, J. Ong, C. Wang, H. Ong, R. Cheng, e D. Ong, «Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation», *Ann. Biomed. Eng.*, vol. 52, fasc. 5, pp. 1115–1118, mag. 2024, doi: 10.1007/s10439-023-03327-6.

[53] T. B. Brown *et al.*, «Language Models are Few-Shot Learners», NeurIPS 2020.

[54] Y. Gao *et al.*, «Retrieval-Augmented Generation for Large Language Models: A Survey», 27 marzo 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[55] L. M. Amugongo, P. Mascheroni, S. G. Brooks, S. Doering, e J. Seidel, «Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review», 11 luglio 2024, *Computer Science and Mathematics*. doi: 10.20944/preprints202407.0876.v1.

[56] D. Edge *et al.*, «From Local to Global: A Graph RAG Approach to Query-Focused Summarization», 19 febbraio 2025, *arXiv*: arXiv:2404.16130. doi: 10.48550/arXiv.2404.16130.

[57] N. Reimers e I. Gurevych, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[58] L. Wang *et al.*, «Text Embeddings by Weakly-Supervised Contrastive Pre-training», 22 febbraio 2024, *arXiv*: arXiv:2212.03533. doi: 10.48550/arXiv.2212.03533.

[59] Y. A. Malkov e D. A. Yashunin, «Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, fasc. 4, pp. 824–836, apr. 2020, doi: 10.1109/TPAMI.2018.2889473.

[60] V. Karpukhin *et al.*, «Dense Passage Retrieval for Open-Domain Question Answering», in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[61] A. Hogan *et al.*, «Knowledge Graphs», *ACM Comput. Surv.*, vol. 54, fasc. 4, pp. 1–37, mag. 2022, doi: 10.1145/3447772.

[62] W. X. Zhao *et al.*, «A Survey of Large Language Models», mar. 2023.

[63] T. Gruber, «A translation approach to portable ontology specifications, Knowledge», vol. 5, fasc. 2, pp. 199–220, 1993.

[64] R. Studer, V. R. Benjamins, e D. Fensel, «Knowledge engineering: Principles and methods», *Data Knowl. Eng.*, vol. 25, fasc. 1–2, pp. 161–197, mar. 1998, doi: 10.1016/S0169-023X(97)00056-6.

[65] «What is SNOMED CT», SNOMED International. Consultato: 19 settembre 2026. [Online]. Disponibile su: https://www.snomed.org/what-is-snomed-ct

[66] R. Angles, «The Property Graph Database Model», 2018.

[67] Donnelly, *SNOMED-CT: The Advanced Terminology and Coding System for eHealth*, vol. 121. Studies in Health Technology and Informatics, 2006.

[68] World Health Organization, «Guidelines for ATC Classification and DDD Assignment». WHO Collaborating Centre for Drug Statistics Methodology, Oslo, 2024.

[69] O. Bodenreider, «The Unified Medical Language System (UMLS): integrating biomedical terminology», *Nucleic Acids Res.*, vol. 32, fasc. 90001, pp. 267D – 270, gen. 2004, doi: 10.1093/nar/gkh061.

[70] «UMLS API Home». Consultato: 13 ottobre 2026. [Online]. Disponibile su: https://documentation.uts.nlm.nih.gov/rest/home.html

[71] MindStick, «Sentence Embeddings - Guide for AI», MindStick. Consultato: 13 ottobre 2026. [Online]. Disponibile su: https://www.mindstick.com/articles/341753/sentence-embeddings-guide-for-ai

[72] T. Mikolov, K. Chen, G. Corrado, e J. Dean, «Efficient Estimation of Word Representations in Vector Space», 7 settembre 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[73] H. Jégou, M. Douze, e C. Schmid, «Product Quantization for Nearest Neighbor Search», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, fasc. 1, pp. 117–128, gen. 2011, doi: 10.1109/TPAMI.2010.57.

[74] E. Norouzi, S. Hertling, e H. Sack, «ConExion: Concept Extraction with Large Language Models», 22 aprile 2025, *arXiv*: arXiv:2504.12915. doi: 10.48550/arXiv.2504.12915.

[75] U. Zaratiana, N. Tomeh, P. Holat, e T. Charnois, «GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer», in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 5364–5376. doi: 10.18653/v1/2024.naacl-long.300.

[76] P. M. Nadkarni, L. Ohno-Machado, e W. W. Chapman, «Natural language processing: an introduction», *J. Am. Med. Inform. Assoc.*, vol. 18, fasc. 5, pp. 544–551, set. 2011, doi: 10.1136/amiajnl-2011-000464.

[77] V. Singh, «Replace or Retrieve Keywords In Documents at Scale», 9 novembre 2017, *arXiv*: arXiv:1711.00046. doi: 10.48550/arXiv.1711.00046.

[78] Levenshtein, «Binary Codes Capable of Correcting Deletions, Insertions, and Reversals», Soviet Physics Doklady, vol. 10, fasc. 8, pp. 707–710, 1966.

[79] Bird, «NLTK: the natural language toolkit», ACL, pp. 69–72, 2006.

[80] R. Sharnagat, «Named Entity Recognition: A Literature Survey».

[81] J. Li, A. Sun, J. Han, e C. Li, «A Survey on Deep Learning for Named Entity Recognition», *IEEE Trans. Knowl. Data Eng.*, vol. 34, fasc. 1, pp. 50–70, gen. 2022, doi: 10.1109/TKDE.2020.2981314.

[82] Y. Lecun, L. Bottou, Y. Bengio, e P. Haffner, «Gradient-based learning applied to document recognition», *Proc. IEEE*, vol. 86, fasc. 11, pp. 2278–2324, nov. 1998, doi: 10.1109/5.726791.

[83] Y. Bengio, P. Simard, e P. Frasconi, «Learning long-term dependencies with gradient descent is difficult», *IEEE Trans. Neural Netw.*, vol. 5, fasc. 2, pp. 157–166, mar. 1994, doi: 10.1109/72.279181.

[84] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney, e A. Lewis, «GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface», EMNLP 2025

[85] J. Johnson, M. Douze, e H. Jegou, «Billion-Scale Similarity Search with GPUs», *IEEE Trans. Big Data*, vol. 7, fasc. 3, pp. 535–547, lug. 2021, doi: 10.1109/TBDATA.2019.2921572.

[86] R. Nogueira e K. Cho, «Passage Re-ranking with BERT», 14 aprile 2020, *arXiv*: arXiv:1901.04085. doi: 10.48550/arXiv.1901.04085.

[87] X. Ma, X. Zhang, R. Pradeep, e J. Lin, «Zero-Shot Listwise Document Reranking with a Large Language Model», 3 maggio 2023, *arXiv*: arXiv:2305.02156. doi: 10.48550/arXiv.2305.02156.

[88] W. Sun *et al.*, «Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents», in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 14918–14937. doi: 10.18653/v1/2023.emnlp-main.923.

[89] E. Tutubalina, Z. Miftahutdinov, V. Muravlev, e A. Shneyderman, «A Comprehensive Evaluation of Biomedical Entity-centric Search», in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Abu Dhabi, UAE: Association for Computational Linguistics, 2022, pp. 596–605. doi: 10.18653/v1/2022.emnlp-industry.61.

[90] C. Ye e C. S. Mitchell, «LLM as Entity Disambiguator for Biomedical Entity-Linking».

[91] F. Gallego, P. Ruas, F. M. Couto, e F. J. Veredas, «Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero- and few-shot scenarios», *Knowl.-Based Syst.*, vol. 314, p. 113211, apr. 2025, doi: 10.1016/j.knosys.2025.113211.

[92] Y. Gu *et al.*, «Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing», *ACM Trans. Comput. Healthc.*, vol. 3, fasc. 1, pp. 1–23, gen. 2022, doi: 10.1145/3458754.

[93] L. Zheng *et al.*, «Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena», 24 dicembre 2023, *arXiv*: arXiv:2306.05685. doi: 10.48550/arXiv.2306.05685.

[94] Van Rijsbergen, *Information Retrieval*, Seconda. Butterworths, 1979.

[95] Powers, «Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation», vol. 2, fasc. 1, pp. 37–63, 2011.

[96] Obsidian Development, *Obsidian*. (2020). [Online]. Disponibile su: https://obsidian.md

[97] «Come funziona Amazon Comprehend Medical - Amazon Comprehend Medical». Consultato: 9 ottobre 2026. [Online]. Disponibile su: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-howitworks.html

[98] P. Bhatia, B. Celikkaya, M. Khalilia, e S. Senthivel, «Comprehend Medical: A Named Entity Recognition and Relationship Extraction Web Service», in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA: IEEE, dic. 2019, pp. 1844–1851. doi: 10.1109/ICMLA.2019.00297.

[99] P. Wehner, C. Piberger, e D. Gohringer, «Using JSON to manage communication between services in the Internet of Things», in *2014 9th International Symposium on Reconfigurable and Communication-Centric Systems-on-Chip (ReCoSoC)*, Montpellier, France: IEEE, mag. 2014, pp. 1–4. doi: 10.1109/ReCoSoC.2014.6861361.

[100] «SNOMED CT linking - Amazon Comprehend Medical». Consultato: 9 ottobre 2026. [Online]. Disponibile su: https://docs.aws.amazon.com/comprehend-medical/latest/dev/ontology-linking-snomed.html