

POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Biomedica

TESI DI LAUREA MAGISTRALE

*VALUTAZIONE DELLA QUALITÀ DELLA REFERTAZIONE CLINICA
NELL'ECOSISTEMA DIGITALE SANITARIO:
IL RUOLO DELL'INTELLIGENZA ARTIFICIALE*



**Politecnico
di Torino**

RELATORI:

Prof./Dott.ssa Balestra Gabriella

Dott. Provini Andrea

LAUREANDA:

Abbiati Sophie

Matricola 330829

ANNO ACCADEMICO 2025-2026

Alla mia famiglia,

Fonte inesauribile d'Amore.

Sommario

CAPITOLO 1 INTRODUZIONE	9
1.1. CONTESTO DELLA TRASFORMAZIONE DIGITALE NELLA SANITÀ	9
1.2. OBIETTIVO DELLA TESI	11
1.3. RILEVANZA E CONTRIBUTO	12
CAPITOLO 2 STATO DELL'ARTE	14
2.1. DEFINIZIONE ED ELEMENTI PRINCIPALI DEL REFERTO CLINICO	14
2.2. TIPOLOGIE DI REFERTO CLINICO	15
2.2.1. REFERTO DI LABORATORIO	16
2.2.2. REFERTO DI DIAGNOSTICA PER IMMAGINI	16
2.2.3. REFERTO DI ANATOMIA PATOLOGICA	17
2.2.4. REFERTO SPECIALISTICO AMBULATORIALE	17
2.3. CARTELLA CLINICA ELETTRONICA (CCE) E IL RUOLO DELLA REFERTAZIONE DIGITALE	19
2.4. NORMATIVE E STANDARD DI RIFERIMENTO	21
2.4.1. CODICE DELL'AMMINISTRAZIONE DIGITALE (CAD)	22
2.4.2. REGOLAMENTO GENERALE SULLA PROTEZIONE DEI DATI (GDPR)	22
2.4.3. REGOLAMENTO SUI DISPOSITIVI MEDICI – MDR	22
2.4.4. ARTIFICIAL INTELLIGENCE ACT – AI ACT	23
2.4.5. CODICE DEONTOLOGICO DEL MEDICO	24
2.4.6. STANDARD INTERNAZIONALI HL7, FHIR E DICOM	24
2.4.7. JOINT COMMISSION INTERNATIONAL	25
2.5. ORGANIZZAZIONE NELLA GESTIONE E CONTROLLO QUALITÀ DEI REFERTI AL CDI	26
CAPITOLO 3 MATERIALI E METODO	28
3.1. CONTESTO PROGETTUALE E OBIETTIVO DELLO STUDIO	28
3.2. DISCOVERY E ANALISI DEL PROCESSO AS-IS	29
3.3. ATTORI COINVOLTI E CRITICITÀ OPERATIVE	30
3.4. PARAMETRI DI VALUTAZIONE DEI REFERTI	32
3.5. PROCESSO ITERATIVO DI CONFRONTO E AFFINAMENTO	33
3.6. DATASET DISPONIBILI E DATASET INCLUSI NELL'ANALISI QUANTITATIVA	35
3.7. PRE-PROCESSING E MATCHING AI-UMANO	37
3.8. NORMALIZZAZIONE DELLE VARIABILI	37

3.9. DEFINIZIONE DEI KPI	38
3.10. METRICHE DI CONCORDANZA TRA AI E VALUTAZIONE UMANA	41
3.11. KPI ORGANIZZATIVI PER L'IDENTIFICAZIONE DEI REFERTI DA REVISIONARE	45
3.12. IMPLEMENTAZIONE DELL'ANALISI IN MATLAB	47
 CAPITOLO 4 RISULTATI	 48
 4.1. REPORT PROGETTUALI E ANDAMENTO GENERALE DELLA CONCORDANZA AI-UMANO	 48
4.2. ANALISI DEL PARAMETRO ACRONIMI	49
4.3. ANALISI DEL PARAMETRO COPY-PASTE	50
4.4. ANALISI DEL PARAMETRO STRUTTURA	52
4.5. ANALISI DEL PARAMETRO QUALITÀ	53
4.6. ANALISI DEI KPI CALCOLATI	54
4.6.1. KPI DESCRITTIVI DEGLI OUTPUT AI	54
4.6.2. KPI AVANZATI DI QUALITÀ DELL'OUTPUT AI	56
4.6.3. CONCORDANZA TRA VALUTAZIONE AI E VALUTAZIONE UMANA	59
4.6.4. KPI ORGANIZZATIVI PER L'IDENTIFICAZIONE DEI REFERTI DA REVISIONARE	62
4.6.5. ANALISI DELLA DIREZIONE DELL'ERRORE	65
4.7. SINTESI DEI RISULTATI PRINCIPALI	66
 CAPITOLO 5 DISCUSSIONE	 67
 CAPITOLO 6 CONCLUSIONI	 71

Indice delle Tabelle

TABELLA 1: CLASSIFICAZIONE RIASSUNTIVA DELLE TIPOLOGIE DI REFERTO CLINICO.	18
TABELLA 2: COMPONENTI PRINCIPALI DEL REFERTO CLINICO.	18
TABELLA 3: ATTORI COINVOLTI NEL CONTROLLO QUALITÀ DEI REFERTI.	27
TABELLA 4: PARAMETRI VALUTATI.	33
TABELLA 5: ELEMENTI PRINCIPALI DEI SEI REPORT.	35
TABELLA 6: KPI DESCRITTIVI DI BASE.	39
TABELLA 7: KPI AVANZATI.	40
TABELLA 8: KPI DI CONCORDANZA TRA AI E VALUTAZIONE UMANA.	42
TABELLA 9: KPI ORGANIZZATIVI.	46
TABELLA 10: PERCENTUALI DI CONCORDANZA TRA VALUTAZIONE AUTOMATICA E UMANA.	49

Indice delle figure

FIGURA 1: DESCRIZIONE DELLE ESIGENZE DELLA DIREZIONE SANITARIA E DEL RESPONSABILE QUALITÀ.	31
FIGURA 2: PRE-PROCESSING DEI REFERTI CLINICI.	33
FIGURA 3: SCHEMA DI PROCESSO ITERATIVO DI AFFINAMENTO DEL MODELLO.	35
FIGURA 4: ANDAMENTO NEL TEMPO DELLA CONCORDANZA AI-UMANO DEL PARAMETRO ACRONIMI.....	50
FIGURA 5: ANDAMENTO NEL TEMPO DELLA CONCORDANZA AI-UMANO DEL PARAMETRO COPIA-INCOLLA.	51
FIGURA 6: ANDAMENTO NEL TEMPO DELLA CONCORDANZA AI-UMANO DEL PARAMETRO STRUTTURA.....	52
FIGURA 7: ANDAMENTO NEL TEMPO DELLA CONCORDANZA AI-UMANO DEL PARAMETRO QUALITÀ ESATTA.....	53
FIGURA 8: CONFRONTO DEGLI INDICATORI PERCENTUALI RILEVATI DALL'ALGORITMO NEI DATASET PRELIMINARE E FINALE DI VALIDAZIONE.	55
FIGURA 9: CONFRONTO DEI PUNTEGGI MEDI NORMALIZZATI RELATIVI ALLA QUALITÀ DEI REFERTI E AL LIVELLO DI COPIA- INCOLLA IMPROPRIO NEI DUE DATASET ANALIZZATI.....	56
FIGURA 10: GLOBAL QUALITY SCORE MEDIO CALCOLATO TRAMITE AI.	58
FIGURA 11: ERROR BURDEN INDEX CALCOLATO TRAMITE AI.....	58
FIGURA 12: REFERTI CONFORMI E QUASI CONFORMI SECONDO L'AI.....	59
FIGURA 13: KPI DELLO SCOSTAMENTO MEDIO TRA VALUTAZIONE AI E VALUTAZIONE UMANA.	61
FIGURA 14: COHEN'S KAPPA E WEIGHTED KAPPA.	62
FIGURA 15: KPI PER L'IDENTIFICAZIONE DEI REFERTI DA REVISIONARE.	64

Abstract

La progressiva digitalizzazione dei sistemi sanitari ha modificato le modalità di produzione, gestione e condivisione della documentazione clinica, evidenziando l'importanza di disporre di documenti sanitari chiari, completi, strutturati e facilmente analizzabili. In questo scenario, il referto clinico assume un ruolo particolarmente rilevante, poiché costituisce uno strumento di comunicazione tra professionisti sanitari, strutture e pazienti e contribuisce alla continuità e alla tracciabilità del percorso di cura. Nonostante la sua importanza, il controllo qualità dei referti è ancora spesso affidato a processi di revisione manuale, onerosi in termini di tempo e difficilmente applicabili in modo sistematico su grandi volumi documentali.

In questo contesto, la presente tesi analizza il possibile impiego di un sistema di intelligenza artificiale basato su Large Language Models come supporto alla valutazione della qualità dei referti poliambulatoriali prodotti presso il Centro Diagnostico Italiano. L'algoritmo è stato sviluppato da ReportAId, una piattaforma che utilizza l'intelligenza artificiale per leggere i referti medici ed estrarre in modo automatico e strutturato informazioni utili per le strutture ospedaliere, garantendo al contempo la protezione dei dati personali. I referti, opportunamente anonimizzati, sono stati valutati dal sistema secondo quattro parametri principali: qualità complessiva, struttura del referto, presenza di acronimi o simboli non consentiti e utilizzo improprio del copia-incolla. Le valutazioni automatiche sono state successivamente confrontate con le corrispondenti valutazioni manuali effettuate da un esperto, al fine di analizzare il grado di concordanza tra sistema AI e giudizio umano.

Il contributo del lavoro consiste nella rielaborazione quantitativa dei dati mediante MATLAB e nella definizione di un framework di Key Performance Indicators finalizzato a descrivere le prestazioni del sistema, la qualità degli output algoritmici e il possibile valore organizzativo del sistema proposto.

Nel set finale di validazione, il sistema ha mostrato un'elevata capacità di individuare i referti classificati dall'esperto come meritevoli di revisione, raggiungendo un recall pari

al 96,92%. La specificità, pari al 19,35%, evidenzia invece una minore capacità di riconoscere correttamente i referti che non richiedono ulteriore controllo, confermando una tendenza prudentiale dell'algoritmo. Il confronto tra AI e valutazione umana ha mostrato risultati particolarmente favorevoli per il parametro relativo al copia-incolla, mentre la qualità complessiva si è confermata la dimensione più complessa da automatizzare, in quanto maggiormente influenzata da criteri interpretativi e contestuali.

I risultati ottenuti suggeriscono che l'intelligenza artificiale possa rappresentare uno strumento utile per il pre-screening e il supporto alla revisione dei referti, contribuendo a rendere il controllo qualità più sistematico, tracciabile e scalabile. Il sistema, tuttavia, non deve essere interpretato come sostitutivo immediato del giudizio umano, che rimane essenziale per la valutazione finale della documentazione clinica, ma come un supporto da integrare nel processo di revisione.

The progressive digitalization of healthcare systems has changed the way clinical documentation is produced, managed and shared, highlighting the importance of clear, complete, structured and easily analyzable healthcare documents. In this scenario, the clinical report plays a particularly relevant role, as it represents a communication tool between healthcare professionals, institutions and patients, and contributes to the continuity and traceability of the care pathway. Despite its relevance, report quality control is still often based on manual review processes, which are time-consuming and difficult to apply systematically to large volumes of documents.

In this context, this thesis analyzes the potential use of an artificial intelligence system based on Large Language Models as a support tool for assessing the quality of outpatient clinical reports produced at Centro Diagnostico Italiano. The algorithm was developed by ReportAId, a platform that uses artificial intelligence to read medical reports and automatically extract structured information useful for healthcare organizations, while ensuring the protection of personal data. The reports, appropriately anonymized, were evaluated by the system according to four main parameters: overall quality, report structure, presence of non-permitted acronyms or symbols, and improper use of copy-

paste. The automatic evaluations were then compared with the corresponding manual assessments performed by an expert, in order to analyze the level of agreement between the AI system and human judgment.

The contribution of this work consists in the quantitative reprocessing of the data using MATLAB and in the definition of a framework of Key Performance Indicators aimed at describing the performance of the system, the quality of the algorithmic outputs and the possible organizational value of the proposed approach.

In the final validation set, the system showed a high ability to identify the reports classified by the expert as requiring revision, achieving a recall of 96.92%. The specificity, equal to 19.35%, instead highlights a lower ability to correctly recognize reports that do not require further review, confirming a cautious tendency of the algorithm. The comparison between AI and human evaluation showed particularly favorable results for the copy-paste parameter, while overall quality was confirmed as the most complex dimension to automate, as it is more strongly influenced by interpretative and contextual criteria.

The results obtained suggest that artificial intelligence may represent a useful tool for pre-screening and supporting report review, helping to make quality control more systematic, traceable and scalable. However, the system should not be interpreted as an immediate replacement for human judgment, which remains essential for the final evaluation of clinical documentation, but rather as a support tool to be integrated into the review process. In perspective, Centro Diagnostico Italiano aims to fully or partially automate the work currently carried out manually by the Healthcare Management Department in the evaluation of report quality

Capitolo 1

Introduzione

1.1. Contesto della trasformazione digitale nella sanità

Negli ultimi anni, il settore sanitario è stato interessato da una progressiva trasformazione digitale, che ha modificato non solo le attività clinico-diagnostiche, ma anche i processi organizzativi, gestionali e documentali delle strutture sanitarie [1]. L'integrazione crescente di tecnologie digitali nei dispositivi medici e nei sistemi informativi sanitari ha favorito una gestione più ampia e articolata dei dati clinici, contribuendo allo sviluppo di percorsi di cura più strutturati, tracciabili e orientati al paziente [2]. Parallelamente, sono cambiate le modalità di raccolta, archiviazione, condivisione e utilizzo delle informazioni sanitarie, con effetti rilevanti sulla continuità assistenziale, sul coordinamento tra professionisti e sull'efficienza complessiva dei processi di cura [3].

In questo scenario, la Cartella Clinica Elettronica rappresenta uno degli strumenti principali della modernizzazione sanitaria. Essa costituisce l'evoluzione digitale della tradizionale cartella clinica cartacea e consente di raccogliere e integrare informazioni provenienti da fonti diverse, tra cui esami diagnostici, referti clinici, parametri vitali, terapie e dati amministrativi [4]. La CCE non può quindi essere considerata soltanto un archivio digitale, ma rappresenta una vera e propria infrastruttura informativa, utile a garantire tracciabilità, accessibilità e continuità del dato clinico, oltre che a supportare l'attività decisionale dei professionisti sanitari [5].

Tuttavia, la semplice disponibilità dei documenti in formato elettronico non è sufficiente, da sola, a determinare un miglioramento effettivo della qualità dei processi clinici. Perché

le informazioni sanitarie possano essere realmente valorizzate, è necessario che la documentazione sia chiara, completa, strutturata e utilizzabile anche in un'ottica di analisi e controllo qualità. In questo senso, la digitalizzazione non riguarda soltanto la conversione del documento da cartaceo a elettronico, ma implica anche la possibilità di rendere i dati clinici più accessibili, confrontabili e misurabili.

Tra i documenti più rilevanti all'interno del percorso assistenziale vi è il referto clinico, che rappresenta uno strumento fondamentale di comunicazione tra professionisti sanitari, strutture e pazienti. Il referto non ha soltanto una funzione descrittiva, ma contribuisce alla trasmissione delle informazioni cliniche, alla continuità della presa in carico e alla documentazione delle decisioni diagnostiche e terapeutiche. Per questo motivo, la sua qualità assume un ruolo centrale sia dal punto di vista clinico sia dal punto di vista organizzativo.

Nonostante la sua importanza, la refertazione clinica si basa ancora spesso su testi liberi e su scelte redazionali dipendenti dal singolo professionista. Questa variabilità può rendere i referti meno uniformi, più difficili da confrontare e più complessi da analizzare in modo sistematico. Inoltre, il controllo qualità dei referti è frequentemente affidato a processi di revisione manuale, che possono risultare onerosi in termini di tempo, dipendenti dall'esperienza del valutatore e difficilmente scalabili su grandi volumi documentali [6].

Da queste criticità emerge l'esigenza di introdurre strumenti in grado di supportare l'analisi della documentazione clinica in modo più oggettivo, sistematico e replicabile. In questa prospettiva, il referto può essere considerato non soltanto come un documento descrittivo, ma anche come un oggetto valutabile secondo criteri definiti, quali chiarezza, completezza, coerenza, struttura e conformità a standard redazionali.

All'interno di questo contesto si inserisce il crescente interesse verso l'impiego dell'intelligenza artificiale e, in particolare, dei Large Language Models (LLM) nell'analisi dei testi clinici. Grazie alla loro capacità di elaborare il linguaggio naturale e di riconoscere informazioni all'interno di documenti complessi, tali modelli possono rappresentare strumenti di supporto alla valutazione della qualità della refertazione clinica. Il loro utilizzo non va interpretato come sostitutivo del giudizio umano, ma come possibile supporto a processi di revisione più strutturati, tracciabili e scalabili.

1.2. Obiettivo della tesi

La ricerca nasce dalla constatazione che, in ambito sanitario, l'introduzione di tecnologie innovative non coincide automaticamente con un miglioramento dei processi o della qualità dei servizi. Per questo motivo, il lavoro non si limita a considerare l'impiego di strumenti di intelligenza artificiale nella refertazione clinica, ma si propone di valutarne le prestazioni attraverso indicatori quantitativi e qualitativi, con l'obiettivo di comprenderne il possibile contributo al controllo qualità dei referti.

Il lavoro prende in esame l'impiego di sistemi di intelligenza artificiale basati su Large Language Models come strumenti di supporto all'analisi automatizzata della refertazione, con l'obiettivo di valutarne le prestazioni e il possibile contributo all'interno di un processo di controllo qualità.

Nel dettaglio, gli obiettivi della ricerca sono:

1. Analizzare le valutazioni prodotte da sistemi di Large Language Models sviluppati dal team ReportAid su referti poliambulatoriali reali, al fine di esaminare specifici indicatori di qualità documentale, quali qualità complessiva, struttura del referto, presenza di acronimi o simboli non consentiti e copia-incolla improprio. L'analisi considera, inoltre, il confronto tra output algoritmico e valutazione umana, con l'obiettivo di comprendere il grado di allineamento del sistema rispetto al giudizio del revisore.
2. Costruire un framework di valutazione basato su Key Performance Indicators quantitativi e qualitativi, finalizzato a descrivere le prestazioni del sistema, la qualità degli output generati e il grado di concordanza con la valutazione umana. Tale framework consente di discutere il potenziale valore organizzativo della soluzione come strumento di supporto al processo di revisione e controllo qualità della refertazione clinica.

La tesi si basa sull'analisi di referti clinici reali redatti dai medici del Centro Diagnostico Italiano, Milano. Tali documenti, opportunamente anonimizzati, vengono valutati nell'ambito del progetto ReportAid. I risultati sono presentati seguendo l'evoluzione cronologica del progetto e, successivamente, approfonditi attraverso una rielaborazione quantitativa autonoma dei dati mediante MATLAB. L'obiettivo finale è discutere in

modo critico le prestazioni osservate, i limiti della soluzione adottata e le possibili prospettive di sviluppo per l'integrazione di strumenti di intelligenza artificiale nei processi di controllo qualità della documentazione clinica.

1.3. Rilevanza e contributo

Il contributo della tesi si sviluppa su tre aspetti principali. Il primo riguarda l'analisi dell'applicazione di un sistema basato su Large Language Models per la valutazione di referti poliambulatoriali. In questo contesto, il progetto ReportAId rappresenta un caso applicativo concreto di utilizzo dell'intelligenza artificiale per l'analisi automatizzata di specifici indicatori di qualità documentale, quali qualità complessiva, struttura del referto, presenza di acronimi o simboli non consentiti e copia-incolla improprio.

Il secondo contributo consiste nella definizione e applicazione di un sistema di Key Performance Indicators, finalizzato a valutare in modo più oggettivo le prestazioni del sistema. I KPI considerati permettono di analizzare la qualità degli output prodotti dall'AI, il grado di concordanza con la valutazione umana e la presenza di eventuali discrepanze tra sistema automatico e revisore. In questo modo, l'introduzione dell'intelligenza artificiale non viene descritta soltanto come innovazione tecnologica, ma viene interpretata attraverso misure utili a comprenderne affidabilità, limiti e potenziale utilità operativa.

Infine, la tesi propone una riflessione sul possibile valore organizzativo dell'intelligenza artificiale nel controllo qualità della refertazione clinica. I risultati ottenuti suggeriscono che strumenti di questo tipo possano supportare una revisione più sistematica e tracciabile dei referti, favorendo l'identificazione dei documenti potenzialmente critici e contribuendo alla standardizzazione dei criteri di valutazione. Tali aspetti devono tuttavia essere interpretati come possibili ricadute applicative, e non come effetti organizzativi direttamente misurati nel presente studio.

Nel complesso, il lavoro intende offrire un contributo metodologico allo studio dell'innovazione digitale in sanità, proponendo un primo framework di valutazione basato su KPI applicato alla refertazione clinica. Tale impostazione potrà essere sviluppata in futuro attraverso l'analisi di campioni più ampi, l'integrazione di dati di

processo e la valutazione diretta dell'impatto operativo della soluzione in condizioni reali di utilizzo.

Capitolo 2

Stato dell'arte

Come accennato in precedenza, nel processo di digitalizzazione sanitaria la refertazione clinica rappresenta un passaggio fondamentale, poiché raccoglie e sintetizza i dati diagnostici e li trasforma in informazioni utili per il medico e per il paziente. Poiché la presente tesi si concentra sul tema della qualità della refertazione clinica, analizziamo in primo luogo la struttura di un referto e le sue caratteristiche principali, facendo poi riferimento all'ecosistema del Centro Diagnostico Italiano, sede della presente tesi. È stata poi effettuata una panoramica delle normative e degli standard di riferimento da rispettare per la redazione di un referto clinico.

2.1. Definizione ed elementi principali del referto clinico

Con il termine *referto* si intende il documento medico-legale mediante il quale il professionista sanitario riporta, in forma scritta, ufficiale e firmata, i risultati di un esame clinico, laboratoristico o strumentale. Dal punto di vista etimologico, la parola “*referto*” deriva dal latino medievale *refertum* (da *referre*, “riportare”), a sottolinearne la funzione di restituzione e condivisione di un'informazione clinica [7].

Il referto non ha soltanto valore descrittivo, ma assume anche una valenza comunicativa e decisionale: rappresenta, infatti, il principale strumento di interscambio di informazioni tra professionisti sanitari e tra medico e paziente, contribuendo alla continuità assistenziale e alla sicurezza delle cure. Inoltre, riveste un ruolo fondamentale sotto il

profilo amministrativo e medico-legale, costituendo un documento ufficiale con valore probatorio e obbligo di conservazione [8][9].

Le caratteristiche fondamentali di un referto clinico comprendono la completezza, intesa come la presenza di tutte le informazioni essenziali, quali i dati identificativi del paziente, la tipologia di esame, la data di esecuzione, i risultati, nonché l'interpretazione e la valutazione clinica. A tale completezza si affianca la chiarezza, garantita dall'impiego di un linguaggio scientifico standardizzato, privo di ambiguità e comprensibile anche per il paziente. È altresì rilevante la tempestività, che si traduce nella disponibilità del referto entro tempi congrui rispetto alla presa in carico del paziente. Infine, il referto possiede un valore probatorio, conferendo al documento validità giuridica e imponendo obblighi di autenticazione, tracciabilità e conservazione secondo le normative vigenti [8][9].

Dal punto di vista ingegneristico, il processo di refertazione clinica può essere interpretato come un flusso strutturato composto da:

- input (dati clinici, esami diagnostici, osservazioni);
- trasformazione (interpretazione e formalizzazione testuale);
- output (referto clinico validato);
- feedback (eventuali revisioni, richieste di chiarimento, integrazioni).

La digitalizzazione ha reso questo processo tracciabile e potenzialmente misurabile. Tuttavia, nella pratica operativa, la qualità dei referti risulta spesso caratterizzata da elevata variabilità strutturale e linguistica, con differenze significative tra professionisti, unità operative e contesti organizzativi. Tale variabilità introduce criticità in termini di standardizzazione, leggibilità, interoperabilità e controllo qualità, rendendo necessario lo sviluppo di strumenti di analisi automatizzata e di metriche oggettive di valutazione.

2.2. Tipologie di referto clinico

I referti clinici possono assumere forme differenti a seconda della tipologia di prestazione sanitaria effettuata. Ciascuna categoria presenta proprie specificità in termini di contenuti, struttura e finalità, pur mantenendo caratteristiche generali di completezza, chiarezza e valore legale. Tra le principali tipologie si distinguono i referti di laboratorio, di

diagnostica per immagini, di anatomia patologica e di prestazioni specialistiche ambulatoriali.

Nell'ambito di questa tesi, particolare attenzione sarà rivolta ai referti poliambulatoriali redatti dai medici del Centro Diagnostico Italiano, essendo complessi e variabili in funzione della prestazione e del dottore che li produce.

Oltre alla valutazione della qualità, in prospettiva futura emerge la possibilità di arricchire i referti con parametri biomedici supportati dall'intelligenza artificiale, permettendo di ottenere informazioni più dettagliate, di analizzare l'andamento dei dati nel tempo e di individuare correlazioni tra esami differenti.

2.2.1. Referto di laboratorio

Il referto di laboratorio riporta i risultati delle analisi chimico-cliniche, microbiologiche, immunologiche o genetiche eseguite su campioni biologici, come sangue, urine, tessuti, etc. Si tratta di un documento altamente standardizzato, in cui i dati sono espressi in forma numerica o testuale, accompagnati da unità di misura e intervalli di riferimento. In questo ambito, la produzione e gestione del referto sono strettamente integrate con i *Laboratory Information System* (LIS), sistemi informativi dedicati alla registrazione, elaborazione e trasmissione dei dati di laboratorio. Tali sistemi contribuiscono ad automatizzare il flusso operativo, ridurre il rischio di errori trascrittivi e garantire una maggiore tracciabilità del processo analitico [11].

2.2.2. Referto di diagnostica per immagini

Il referto di diagnostica per immagini accompagna gli esami radiologici e di imaging, comprendendo tra questi la radiografia, la tomografia computerizzata (TC), la risonanza magnetica (RM), l'ecografia e le indagini di medicina nucleare. È redatto dal medico radiologo e descrive in modo dettagliato le osservazioni morfologiche e funzionali rilevate nelle immagini, concludendosi generalmente con un giudizio diagnostico o con ipotesi differenziali.

Negli ultimi anni si è progressivamente diffusa la refertazione strutturata, che prevede l'utilizzo di modelli predefiniti e terminologie codificate, con l'obiettivo di ridurre

l'ambiguità interpretativa, aumentare l'uniformità dei documenti e favorire la successiva elaborazione automatizzata dei dati [12].

Un esempio rilevante è rappresentato da RadLex, lessico radiologico sviluppato dalla Radiological Society of North America (RSNA), finalizzato a fornire una terminologia condivisa per la refertazione, il supporto decisionale, il data mining e la ricerca in ambito radiologico [13]. L'impiego di terminologie standardizzate consente di migliorare la comunicazione tra radiologo e medico richiedente, favorendo al contempo l'interoperabilità semantica delle informazioni.

2.2.3. Referto di anatomia patologica

Il referto di anatomia patologica riporta i risultati delle analisi istologiche, citologiche o autoptiche su campioni di tessuto o cellule. È redatto dallo specialista anatomo-patologo che spesso fornisce la conferma definitiva della natura benigna o maligna di una lesione. Il documento descrive le caratteristiche macroscopiche e microscopiche del campione, l'eventuale presenza di marker immunoistochimici e la diagnosi conclusiva. La standardizzazione e la tracciabilità del referto anatomo-patologico sono essenziali per garantire la qualità diagnostica, la sicurezza del paziente e l'interoperabilità con registri tumori e sistemi informativi ospedalieri [14].

2.2.4. Referto specialistico ambulatoriale

I referti specialistici ambulatoriali sono redatti da medici specialisti al termine di una visita o di una prestazione ambulatoriale (ad esempio cardiologica, dermatologica, ORL, ortopedica). Contengono la descrizione dell'anamnesi, l'esame obiettivo, i risultati di eventuali test diagnostici, la valutazione clinica e le indicazioni terapeutiche o di follow-up. Questi referti hanno una struttura più discorsiva rispetto a quelli di laboratorio e d'imaging, integrando dati clinici e osservazioni dello specialista. Rivestono un ruolo cruciale nella continuità assistenziale, poiché permettono la condivisione delle informazioni cliniche tra il medico di base, gli specialisti e le strutture ospedaliere, diventando il documento principale per pianificare i successivi interventi diagnostici e terapeutici [14]. La loro digitalizzazione e integrazione nei sistemi informativi sanitari regionali costituiscono uno degli obiettivi chiave della sanità digitale italiana [15][16].

Tabella 1: Classificazione riassuntiva delle tipologie di referto clinico.

TIPOLOGIA	CONTENUTO PRINCIPALE	STRUTTURA
Referto Di Laboratorio	Risultati di analisi cliniche (esami ematici, urinari, biochimici)	Presentazione tabellare con valori numerici o qualitativi e intervalli di riferimento
Referto Di Diagnostica Per Immagini	Descrizione di reperti da indagini strumentali (radiografie, ecografie, TAC, RMN)	Testo descrittivo con riferimento a immagini
Referti Di Anatomia Patologica	Analisi di campioni biologici (es. biopsie, citologie)	Referto dettagliato con classificazioni codificate, descrizione macro e micro e diagnosi
Referti Specialistici Ambulatoriali	Valutazioni cliniche afferenti a diversi ambiti specialistici (es. cardiologia, ortopedia, dermatologia)	Testo narrativo strutturati con indicazioni diagnostico-terapeutiche ed eventuali allegati strumentali

I referti delle prestazioni ambulatoriali presentano differenze tra le varie specialità, in particolare per quanto concerne l'anamnesi specialistica e l'esame obiettivo. Tutti i referti comprendono, però, indipendentemente dalla specialità, i settori presentati nella *Tabella 2* sottostante.

Tabella 2: Componenti principali del referto clinico.

VALUTAZIONE	AMBITO / ETÀ	MODALITÀ	CRITERI / NOTE
Rischio Cadute	Tutti i pazienti	Compilata dallo specialista durante la prima visita	Basata sulla scala di Conley modificata; score disponibile nella cartella del paziente
Qualità Della Vita	Pazienti >74 anni (64 anni per visite geriatriche)	Questionario obbligatorio	Valutazione del cammino, autonomia nelle attività quotidiane, dolore, ansia e depressione; se lo score è significativo, indicare consigli o terapia

<i>Dolore</i>	Maggior parte delle visite specialistiche	Scaturisce in caso di dolore riferito >4	Valutazione entità dolore e impatto sulla vita quotidiana tramite scala VAS; se lo score è significativo, indicare consigli o terapie; non presente se compilato il questionario qualità della vita
<i>Valutazione Pediatrica</i>	Pazienti <18 anni; ambito internistico	Questionario obbligatorio	Valutazione vissuto familiare, scolastico e sviluppo psicofisico; se lo score è significativo, indicare consigli o terapie
<i>Masticazione</i>	Visite internistiche, otorinolaringoiatriche, geriatrie, neurologiche	Questionario obbligatorio	Se rilevate difficoltà nella masticazione o deglutizione, indicare consigli o terapie
<i>Udito</i>	Visite internistiche, otorinolaringoiatriche, geriatrie, neurologiche	Questionario obbligatorio	Se rilevata ipoacusia, indicare consigli o terapie
<i>Vista</i>	Visite internistiche, cardiologiche, geriatrie	Questionario obbligatorio	Se rilevati difetti visivi, indicare consigli o terapie

2.3. Cartella Clinica Elettronica (CCE) e il ruolo della refertazione digitale

Poiché l'obiettivo della presente tesi è analizzare il controllo della qualità della refertazione clinica, con particolare attenzione ai fattori tecnologici e organizzativi che la influenzano, si ritiene necessario approfondire le caratteristiche della Cartella Clinica Elettronica.

La CCE rappresenta una delle innovazioni centrali nel processo di digitalizzazione della sanità. Essa costituisce un sistema informatico progettato per la raccolta, la gestione, l'archiviazione e la condivisione dei dati clinici di un paziente in formato digitale, sostituendo progressivamente la tradizionale documentazione cartacea [16]. La CCE consente di integrare in un unico ambiente virtuale informazioni cliniche, diagnostiche e amministrative, offrendo una visione completa e coerente del percorso di cura del

paziente. In particolare, essa si configura come un applicativo di produzione, ampiamente utilizzato per la creazione e la gestione dei referti ambulatoriali.

È importante sottolineare che la Cartella Clinica Elettronica non va confusa con la semplice redazione elettronica dei referti tramite template o strumenti di produttività individuale (ad esempio Word o Excel). Questi ultimi, infatti, permettono solo la creazione digitale del documento, senza garantire l'integrazione, la tracciabilità, la storicizzazione dei dati, né il collegamento automatico con altri sistemi clinici o dispositivi medicali. La CCE, al contrario, rappresenta un sistema strutturato e centralizzato, capace di raccogliere, gestire e rendere disponibili informazioni cliniche in modo coerente, sicuro e condiviso tra più reparti o strutture.

La Cartella Clinica Elettronica non si limita a essere un archivio statico di documenti: rappresenta un sistema attivo di supporto alla pratica clinica. Oltre a garantire la tracciabilità e la storicizzazione di ogni prestazione, può fornire funzionalità di supporto decisionale, come l'attivazione automatica di avvisi in presenza di interazioni farmacologiche o deviazioni dai protocolli terapeutici. Grazie all'accesso controllato e profilato, le informazioni possono essere consultate da professionisti autorizzati anche in modalità remota, assicurando la continuità delle cure tra reparti e strutture sanitarie diverse. Parallelamente, i dati archiviati nella CCE costituiscono una risorsa di grande valore per la ricerca clinica e epidemiologica, consentendo analisi statistiche e studi di outcome utili a migliorare la qualità dei servizi sanitari [16][17][18].

L'adozione di una CCE comporta anche scelte strategiche legate alle modalità di implementazione del sistema informatico. In particolare, le organizzazioni sanitarie possono orientarsi verso due approcci principali: lo sviluppo di una soluzione personalizzata (*make*), oppure l'acquisizione di una piattaforma commerciale già esistente (*buy*). L'approccio *make* consente un'elevata personalizzazione dei flussi clinici e una maggiore flessibilità nell'integrazione con sistemi e dispositivi già presenti, ma richiede investimenti significativi in termini di risorse, competenze e tempi di sviluppo. Al contrario, l'approccio *buy* permette una più rapida implementazione e l'utilizzo di soluzioni già validate, a fronte di possibili limitazioni in termini di adattabilità, interoperabilità e dipendenza dal fornitore. La scelta tra *make* e *buy* risulta strettamente legata agli obiettivi dell'organizzazione sanitaria, al livello di digitalizzazione già

raggiunto e alle performance attese della CCE, in particolare per quanto riguarda l'integrazione dei dati clinici e dei dati provenienti dai dispositivi medicali.

L'adozione della Cartella Clinica Elettronica (CCE) comporta benefici significativi: la digitalizzazione consente di ridurre gli errori medici legati a scarsa leggibilità o dispersione delle informazioni, velocizza la gestione dei referti e permette un accesso immediato e condiviso ai dati del paziente. Inoltre, la struttura digitale del sistema favorisce l'integrazione con tecnologie basate su intelligenza artificiale e machine learning, aprendo la strada a nuovi strumenti di supporto alla diagnosi e alla refertazione automatizzata [16][18].

Nonostante questi vantaggi, l'implementazione della Cartella Clinica Elettronica presenta diverse criticità. La piena interoperabilità tra sistemi eterogenei rimane una delle principali sfide, a causa dell'uso non uniforme degli standard di comunicazione. La sicurezza e la protezione dei dati personali rappresentano un ulteriore aspetto critico che richiede conformità al Regolamento Europeo 2016/679 (GDPR) e alle normative nazionali in materia di privacy [19]. A ciò si aggiungono le resistenze organizzative e culturali del personale sanitario, spesso chiamato ad adattarsi a nuovi strumenti digitali, e i costi elevati legati all'implementazione e alla manutenzione dei sistemi informativi [17][18].

2.4. Normative e standard di riferimento

La gestione dei dati sanitari e la refertazione clinica sono regolate da un insieme di normative nazionali e internazionali, affiancate da standard tecnici che ne supportano l'implementazione operativa. Le normative – come il Codice dell'Amministrazione Digitale (CAD), il Regolamento Generale sulla Protezione dei Dati (GDPR), il Regolamento sui Dispositivi Medici (MDR), l'Artificial Intelligence Act e il Codice Deontologico del Medico – definiscono obblighi legali, principi etici e requisiti per garantire sicurezza, privacy e correttezza nella gestione delle informazioni cliniche. A queste si affiancano gli standard internazionali, quali HL7, FHIR e DICOM, che forniscono specifiche tecniche per far sì che ci sia interoperabilità, uniformità e qualità nello scambio e nella rappresentazione dei dati sanitari.

2.4.1. Codice dell'Amministrazione Digitale (CAD)

Il CAD (D.Lgs. 7 marzo 2005, n. 82 e successive modifiche) definisce i principi e le regole per la digitalizzazione della pubblica amministrazione, con particolare attenzione alla gestione dei documenti elettronici, alla conservazione digitale e all'interoperabilità tra sistemi informativi [20]. Nel contesto sanitario, il CAD stabilisce le regole per la digitalizzazione delle cartelle cliniche e dei referti, la loro firma elettronica qualificata, la conservazione sostitutiva a norma e l'interscambio sicuro tra strutture sanitarie [20][21].

2.4.2. Regolamento Generale sulla Protezione dei Dati (GDPR)

Il GDPR (Regolamento UE 2016/679) disciplina il trattamento dei dati personali, inclusi quelli relativi alla salute, stabilendo obblighi stringenti in materia di privacy, sicurezza e responsabilizzazione dei titolari del trattamento [22]. Nel contesto sanitario, esso richiede che i dati siano trattati nel rispetto dei principi di liceità, minimizzazione, riservatezza e protezione, garantendo adeguate misure di sicurezza, tracciabilità delle operazioni e tutela dei diritti dell'interessato. Tra questi rientrano il diritto di accesso, rettifica e, nei casi previsti, portabilità dei dati [22]. Per i trattamenti necessari all'erogazione delle prestazioni sanitarie, il consenso dell'interessato non costituisce di norma la base giuridica richiesta, fermo restando l'obbligo di fornire un'informativa chiara e completa sul trattamento dei dati [23].

2.4.3. Regolamento sui Dispositivi Medici – MDR

Il Regolamento (UE) 2017/745 sui dispositivi medici, comunemente indicato come Medical Device Regulation (MDR), disciplina l'immissione sul mercato, la messa a disposizione e la sorveglianza dei dispositivi medici all'interno dell'Unione Europea. Il suo obiettivo è garantire elevati standard di sicurezza, prestazione e qualità, rafforzando i requisiti di valutazione clinica, gestione del rischio, tracciabilità e sorveglianza post-commercializzazione dei prodotti destinati all'impiego sanitario [24].

Nel contesto della sanità digitale, il MDR assume particolare rilievo anche per i software medicali. Un software può infatti rientrare nella definizione di dispositivo medico qualora sia destinato dal fabbricante a finalità mediche specifiche, come diagnosi, prevenzione, monitoraggio, previsione, prognosi, trattamento o attenuazione di una malattia [24][25].

Con riferimento al tema della presente tesi, il MDR costituisce un quadro normativo rilevante nel momento in cui strumenti di intelligenza artificiale o sistemi software vengono integrati nei processi clinici con una funzione che incide direttamente sulla valutazione diagnostica, sulla gestione del paziente o sulla produzione di informazioni clinicamente significative. In tali casi, il sistema deve essere progettato e validato secondo criteri stringenti di affidabilità, sicurezza e controllo del rischio, affinché il suo utilizzo non comprometta la tutela del paziente e la qualità dell'assistenza [24][25].

2.4.4. Artificial Intelligence Act – AI Act

L'Artificial Intelligence Act, introdotto con il Regolamento (UE) 2024/1689, rappresenta il primo quadro normativo organico dell'Unione Europea dedicato all'intelligenza artificiale. Il regolamento adotta un approccio risk-based, classificando i sistemi di AI in funzione del livello di rischio che possono generare per la salute, la sicurezza e i diritti fondamentali delle persone [26].

In ambito sanitario, l'AI Act assume particolare importanza perché i sistemi di intelligenza artificiale destinati a finalità mediche possono rientrare nella categoria dei sistemi ad alto rischio. Ciò riguarda, in particolare, i software basati su AI destinati a uso medico o impiegati come componenti di sicurezza di prodotti regolati dalla normativa europea sui dispositivi medici [26][27].

Per tali sistemi, l'AI Act prevede requisiti specifici riguardanti la gestione del rischio, la qualità e la governance dei dati, la documentazione tecnica, la registrazione delle operazioni, la trasparenza verso gli utilizzatori, la supervisione umana, nonché l'accuratezza, la robustezza e la cybersicurezza del sistema [26][27].

Nel quadro della presente tesi, l'AI Act costituisce un riferimento normativo centrale per contestualizzare l'adozione di strumenti di intelligenza artificiale nei processi di valutazione e controllo della qualità dei referti. Anche qualora un sistema non operi in autonomia decisionale, il suo utilizzo in un contesto sanitario richiede attenzione alla trasparenza del funzionamento, alla supervisione da parte di professionisti qualificati e alla valutazione dei possibili rischi connessi a errori, bias o interpretazioni non corrette delle informazioni cliniche [26][27].

2.4.5. Codice Deontologico del Medico

Il Codice di Deontologia Medica definisce i principi etici e professionali che regolano l'esercizio della professione sanitaria, includendo disposizioni rilevanti in materia di documentazione clinica, informazione al paziente, responsabilità professionale e riservatezza dei dati [28].

In relazione alla refertazione, il medico è tenuto a produrre documentazione veritiera, chiara, completa e coerente con l'atto sanitario compiuto. La qualità del referto non riguarda quindi soltanto l'accuratezza tecnica del contenuto, ma anche la sua capacità di rappresentare in modo corretto e comprensibile l'attività clinica svolta e le informazioni necessarie alla continuità assistenziale [28].

2.4.6. Standard internazionali HL7, FHIR e DICOM

Gli standard internazionali adottati garantiscono l'interoperabilità dei dati clinici tra sistemi e dispositivi eterogenei, definendo modalità condivise per la strutturazione, la codifica e lo scambio delle informazioni sanitarie. Tra i principali standard di riferimento si annoverano:

- HL7 CDA (Clinical Document Architecture): definisce il formato dei documenti clinici elettronici, incluso il referto, per favorire lo scambio strutturato di informazioni [29].
- FHIR (Fast Healthcare Interoperability Resources): evoluzione moderna di HL7, basata su risorse modulari e protocolli web, semplifica la condivisione dei dati clinici tra applicazioni eterogenee [30].
- DICOM (Digital Imaging and Communications in Medicine): standard per la gestione, archiviazione e trasmissione delle immagini mediche, garantendo la compatibilità dei sistemi di diagnostica per immagini [31].

L'adozione di questi standard è essenziale per assicurare che referti e immagini diagnostiche siano scambiabili in modo sicuro e interpretabile, favorendo la continuità assistenziale e la qualità delle cure del paziente [29][30][31].

Tuttavia, questo livello di interoperabilità non è sempre pienamente raggiungibile. Le ragioni risiedono principalmente nella frammentazione dei sistemi informativi sanitari,

nella coesistenza di dispositivi medicali eterogenei e spesso basati su protocolli proprietari, nonché nella variabilità con cui le diverse strutture implementano gli standard internazionali [32][33]. A ciò si aggiungono limiti tecnici e organizzativi, come la mancata armonizzazione dei formati dati, la differente maturità digitale delle aziende sanitarie e la complessità dei processi di integrazione tra infrastrutture cliniche, software e medical devices [32][33].

Questi fattori rendono complessa l'integrazione fluida dei dati e possono ostacolare la piena interoperabilità tra dispositivi medicali e sistemi digitali, quali la Cartella Clinica Elettronica, evidenziando così la necessità di ulteriori interventi normativi, tecnologici e organizzativi.

2.4.7. Joint Commission International

Tra gli standard di qualità di riferimento utilizzati presso il Centro Diagnostico Italiano si inserisce l'adozione dei criteri definiti dalla Joint Commission International (JCI), organismo internazionale indipendente che si occupa della definizione e della diffusione di standard per l'accreditamento delle strutture sanitarie a livello globale [34][35].

La JCI ha l'obiettivo di promuovere il miglioramento continuo della qualità dell'assistenza e della sicurezza del paziente attraverso modelli organizzativi e clinici basati su standard misurabili e verificabili. Gli standard JCI si articolano in diverse aree tematiche, tra cui la gestione e la sicurezza del paziente, la continuità delle cure, la governance clinica, la gestione delle informazioni e la qualità della documentazione sanitaria [34].

In questo contesto, la refertazione clinica assume un ruolo centrale, in quanto rappresenta uno strumento fondamentale per la comunicazione tra professionisti sanitari e per la tracciabilità del percorso diagnostico-terapeutico del paziente. Una documentazione accurata, completa e tempestiva contribuisce infatti a ridurre la variabilità dei processi, minimizzare il rischio di errori e sostenere un approccio sistematico alla qualità e alla sicurezza delle cure.

L'adozione di tali standard contribuisce quindi a strutturare la refertazione clinica secondo criteri condivisi e verificabili, favorendo un approccio orientato al miglioramento continuo e all'affidabilità della documentazione sanitaria. Il Centro

Diagnostico Italiano ha rinnovato nel 2025, per la settima volta, il proprio accreditamento Joint Commission International, confermando l'attenzione dell'organizzazione verso la qualità e la sicurezza dei processi assistenziali [35].

2.5. Organizzazione nella gestione e controllo qualità dei referti al CDI

Analizzando più nel dettaglio il network del Centro Diagnostico Italiano (CDI), sede della presente tesi, emerge come la redazione dei referti delle visite specialistiche sia regolata da procedure interne che definiscono con precisione ruoli, responsabilità e modalità operative. L'obiettivo di tali procedure è garantire una gestione uniforme e sicura del referto in tutte le sue fasi — dalla compilazione alla modifica, dall'archiviazione alla successiva consultazione — assicurando la tracciabilità delle operazioni, la protezione dei dati e la conformità alle normative vigenti in materia di sicurezza e conservazione dei documenti sanitari.

La responsabilità generale dell'attuazione e del controllo delle politiche di refertazione è attribuita alla Direzione Sanitaria, che ne verifica la corretta applicazione. Il medico specialista è responsabile della redazione del referto in conformità alle linee guida interne e alle indicazioni del CDI. Egli è inoltre responsabile della cura del paziente e tenuto a refertare all'interno del Sistema di Refertazione poliambulatoriale, accedendo tramite credenziali personali che assicurano la corretta identificazione dell'autore e la tutela della sicurezza dei dati clinici.

La funzione informatica aziendale si occupa della conservazione dei referti digitali secondo le misure previste dal Piano della Sicurezza dei Dati, garantendone l'integrità e la disponibilità nel tempo. La direzione del poliambulatorio ha invece il compito di formare i medici all'utilizzo del sistema di refertazione, assicurando un uso appropriato e standardizzato della piattaforma. Infine, la direzione amministrativa è incaricata dell'archiviazione dei referti che ancora vengono redatti in formato cartaceo (es. referti strumentali ancora non interfacciati in cartella), nel rispetto dei tempi e delle modalità stabilite dalle normative vigenti e dalle procedure interne, tra cui la Procedura di conservazione della cartella clinica e del referto.

Per quanto riguarda il processo di controllo di qualità, due sono gli attori principali coinvolti: il Direttore Sanitario e il Responsabile Qualità. Il Direttore Sanitario ha il compito di controllare manualmente i referti medici trimestralmente e di fornire una valutazione per ciascun documento analizzato. Accanto a questa figura, il Quality Manager svolge un ruolo cruciale nella preparazione, configurazione e validazione dei dati, nonché nel monitoraggio degli indicatori di processo.

Tabella 3: Attori coinvolti nel controllo qualità dei referti.

<i>FIGURA</i>	<i>COMPITI SVOLTI</i>
<i>Direttore Sanitario</i>	▫ Valutazione manuale della qualità dei referti clinici
<i>Responsabile Qualità</i>	▫ Preparazione, configurazione e validazione dei dati ▫ Monitoraggio degli indicatori di processo (KPI)

Il Direttore Sanitario del CDI è attualmente tenuto a valutare manualmente la qualità di centinaia di referti con cadenza trimestrale, con un conseguente significativo dispendio di tempo e risorse. Tale attività si caratterizza per essere altamente ripetitiva, soggettiva e onerosa sotto il profilo operativo.

Inoltre, il sistema di feedback attualmente in uso risulta inefficace nel determinare un miglioramento dei comportamenti clinici, in quanto le segnalazioni di errore vengono fornite con un ritardo tale da non incidere in modo tempestivo sulle pratiche dei medici coinvolti.

Da qui l'idea di implementare un progetto che utilizzi l'intelligenza artificiale in modo tale da sostituire in parte o completamente il lavoro svolto attualmente dall'esperto del CDI. Lo sviluppo del progetto è presentato nei capitoli successivi.

Capitolo 3

Materiali e Metodi

3.1. Contesto progettuale e obiettivo dello studio

Il presente studio si inserisce nel processo di digitalizzazione della refertazione clinica presso il Centro Diagnostico Italiano, con particolare attenzione alla valutazione della qualità dei referti poliambulatoriali. In questo contesto, il controllo qualità del referto è un'attività fondamentale, poiché tale documento rappresenta uno dei principali strumenti di comunicazione tra medico, struttura sanitaria e paziente.

Il progetto ha previsto l'utilizzo di un sistema di intelligenza artificiale basato su LLM, sviluppato dal gruppo ReportAId, per supportare la revisione dei referti clinici. Il sistema analizza i documenti e assegna una valutazione automatica ad alcuni aspetti della qualità del referto, come la Struttura, la Qualità del testo, la presenza di Acronimi non consentiti e l'eventuale uso improprio del Copia-Incolla. Queste valutazioni automatiche sono state poi confrontate con quelle effettuate manualmente dalla Direzione Sanitaria del CDI.

La presente tesi si concentra sull'analisi quantitativa dei risultati prodotti dall'AI. In particolare, gli output dell'algoritmo e le valutazioni effettuate manualmente da esperti sono stati confrontati ed elaborati attraverso un framework di Key Performance Indicators. Questi indicatori permettono di descrivere la qualità dei referti valutati dall'AI, misurare il grado di accordo con il giudizio umano e valutare il possibile ruolo del sistema come supporto al processo di controllo qualità.

L'elaborazione è stata svolta in MATLAB, utilizzato per importare i dati, allineare le valutazioni automatiche e manuali, convertire le variabili in formato numerico e calcolare

gli indicatori. In questo modo, i risultati prodotti dall'AI sono stati trasformati in misure sintetiche e confrontabili, utili per interpretare il comportamento del sistema e individuare eventuali differenze rispetto alla valutazione umana.

3.2. Discovery e analisi del processo AS-IS

La fase di *Discovery* ha rappresentato il punto di partenza dell'analisi del processo di refertazione e revisione. L'obiettivo principale è stato comprendere il contesto operativo reale in cui il sistema di intelligenza artificiale avrebbe dovuto essere introdotto, partendo dall'osservazione delle attività svolte dagli utenti coinvolti e dalle criticità emerse nel flusso di lavoro esistente.

L'approccio adottato è stato di tipo *bottom-up*: l'analisi non è partita da requisiti teorici predefiniti, ma dall'esperienza concreta di medici, revisori e figure coinvolte nel controllo qualità. Sono state considerate le modalità con cui i referti vengono compilati, selezionati, analizzati e revisionati, nonché le principali difficoltà operative riscontrate nella gestione quotidiana del processo.

In questa fase è stata sviluppata una *User Journey Map*, finalizzata a rappresentare in modo strutturato le principali fasi del processo: compilazione del referto, selezione del campione, analisi dei referti clinici, analisi dei dati e implementazione dei feedback. Per ciascuna fase sono stati individuati gli attori coinvolti, le attività svolte e i principali punti di contatto del processo. A partire da tale mappatura, è stata condotta un'analisi documentale volta a raccogliere e organizzare le informazioni disponibili sui processi di controllo qualità, sulle linee guida interne e sui criteri di valutazione dei referti.

Parallelamente, sono stati organizzati incontri di aggiornamento con gli stakeholder coinvolti, in particolare con la Direzione Sanitaria e il team Qualità del CDI, al fine di condividere le evidenze emerse e raccogliere indicazioni operative utili per le fasi successive del progetto. Questa fase ha permesso di costruire una comprensione più completa del processo AS-IS, evidenziando le principali criticità e le opportunità di miglioramento. Le informazioni raccolte sono state utilizzate come base per definire i requisiti funzionali e operativi della soluzione TO-BE.

3.3. Attori coinvolti e criticità operative

Gli attori principali coinvolti nel processo di valutazione della qualità dei referti clinici sono la Direzione Sanitaria e il Responsabile Qualità.

La Direttrice Sanitaria rappresenta la figura responsabile della valutazione manuale della qualità dei referti. Tale attività richiede la revisione periodica di un numero elevato di documenti e comporta un significativo carico operativo. Nel tempo, la Direzione Sanitaria ha sviluppato una conoscenza approfondita delle principali tipologie di errore presenti nei referti, che costituisce un elemento fondamentale per il processo di controllo qualità. Tuttavia, la revisione manuale risulta caratterizzata da tempistiche prolungate, elevato carico cognitivo e presenza di attività ripetitive a basso valore aggiunto, come la verifica di acronimi, simboli non consentiti o casi di copia-incolla improprio.

Il Responsabile Qualità è invece coinvolto nella preparazione dei dati, nel monitoraggio degli indicatori e nella verifica della conformità agli standard interni ed esterni. Nel processo tradizionale emergono alcune criticità rilevanti, tra cui la limitata numerosità del campione analizzato, il ritardo temporale nella disponibilità dei dati e la forte componente manuale del flusso di lavoro. Inoltre, la valutazione manuale può essere influenzata da elementi di soggettività e da una variabilità interpretativa tra valutatori.

In questo scenario, l'introduzione di un sistema AI non ha l'obiettivo iniziale di sostituire direttamente il giudizio umano, ma di supportare il processo di revisione, aumentandone la scalabilità, la tracciabilità e la tempestività. Il sistema può infatti contribuire a identificare in modo preliminare i referti potenzialmente critici, orientando l'attenzione del revisore verso i documenti che richiedono maggiore approfondimento. Inoltre, l'impiego di LLM per la valutazione della qualità dei referti consente di ridurre significativamente i tempi necessari rispetto alla revisione manuale. Mentre quest'ultima dipende da fattori legati al singolo revisore, quali esperienza, carico di lavoro e velocità di analisi, la valutazione automatizzata può essere eseguita in tempi brevi e sostanzialmente costanti: attualmente in media sono necessari circa 30 secondi per valutare un referto, quindi intorno ai 50 minuti, poco meno di 1 ora, per 100 referti.

Ciò permette di disporre di una stima temporale più oggettiva e prevedibile, riducendo la variabilità inevitabilmente associata all'intervento umano. L'obiettivo futuro è ridurre progressivamente l'intervento manuale, fino ad automatizzare completamente tali attività mediante l'intelligenza artificiale.

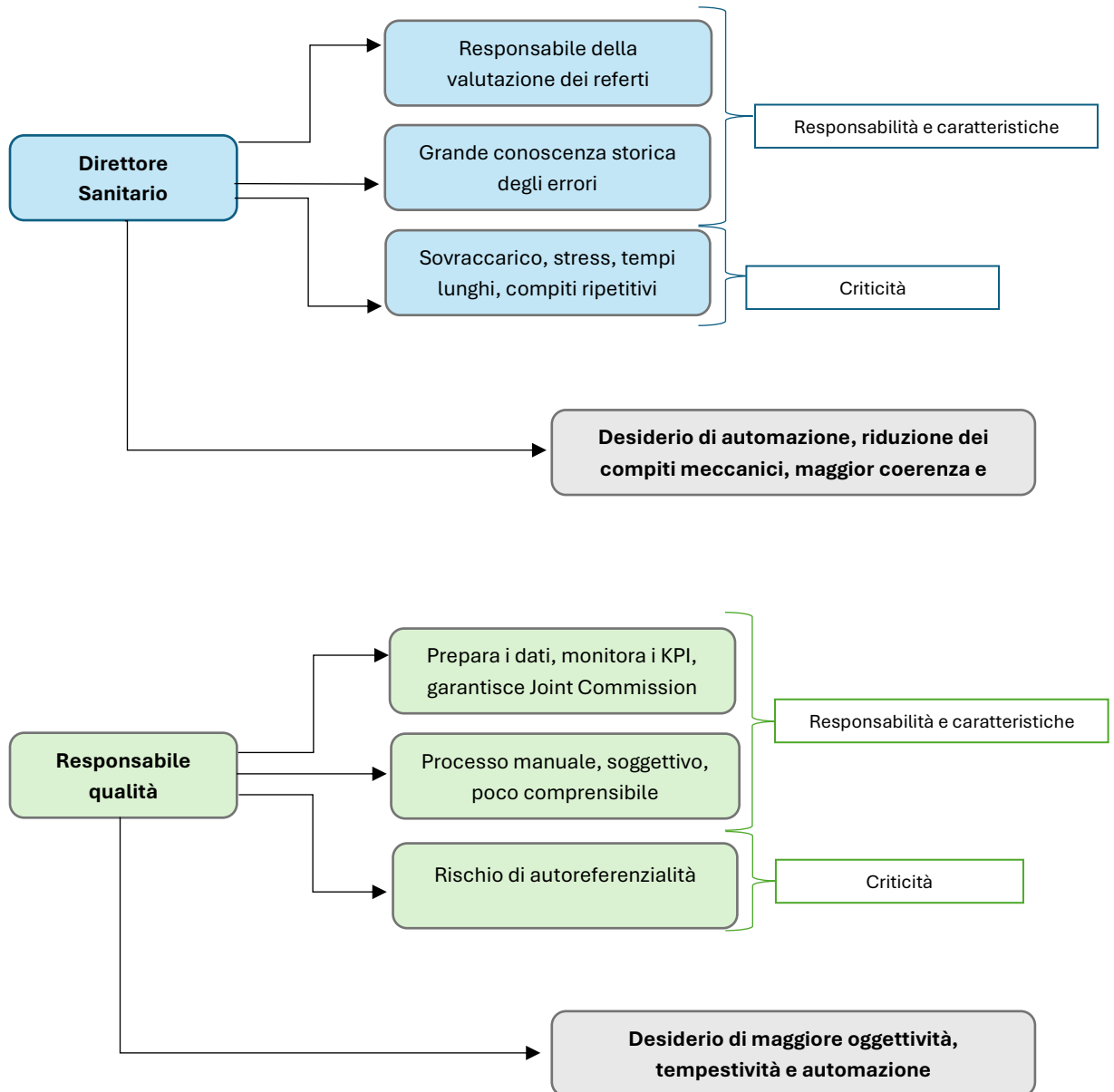


Figura 1: Descrizione delle esigenze della Direzione Sanitaria e del Responsabile Qualità.

3.4. Parametri di valutazione dei referti

La valutazione della qualità dei referti si basa su un insieme di parametri definiti dal CDI e successivamente tradotti in criteri operativi utilizzabili dal sistema AI. I principali parametri considerati sono:

- qualità complessiva del referto;
- presenza di acronimi o simboli non consentiti;
- presenza di copia-incolla improprio;
- correttezza della struttura del referto;
- comunicazione complessiva.

Ai fini dell'analisi quantitativa sviluppata nella presente tesi, sono stati considerati quattro parametri principali: Qualità, Acronimi, Copia-Incolla e Struttura. La comunicazione complessiva, pur rilevante nella fase iniziale di definizione dei criteri, è stata ricondotta al parametro più ampio della qualità complessiva, poiché è molto simile ad essa.

Il parametro Qualità valuta la correttezza complessiva del testo, considerando aspetti quali chiarezza, completezza, coerenza e appropriatezza delle informazioni riportate. È espresso su scala ordinale da 1 a 4, dove valori più elevati indicano una migliore qualità del referto.

Il parametro Acronimi valuta la presenza di sigle, abbreviazioni o simboli non consentiti secondo le linee guida interne del CDI. È trattato come variabile binaria, in cui la presenza di acronimi non consentiti rappresenta una criticità.

Il parametro Copia-Incolla valuta la presenza di contenuti duplicati, ripetizioni anomale o elementi fuori contesto riconducibili a un uso improprio del copia-incolla. È espresso su scala da 0 a 2, dove 0 indica assenza di criticità e valori superiori indicano una criticità crescente.

Il parametro Struttura valuta la corretta organizzazione del referto, con particolare riferimento alla presenza e alla completezza delle sezioni obbligatorie e alla collocazione appropriata delle informazioni cliniche. È trattato come variabile binaria, distinguendo tra struttura adeguata e struttura non adeguata.

Tabella 4: Parametri valutati.

PARAMETRO	DESCRIZIONE	VALORI
<i>Acronimi</i>	Verifica l'uso corretto di abbreviazioni mediche e segnala eventuali termini non consentiti al CDI	SI (presenza di acronimi non consentiti) oppure NO (assenza)
<i>Copia-Incolla</i>	Rileva contenuti duplicati o fuori contesto dovuti a copia-incolla improprio	Punteggi 0 (nessuna ripetizione), 1 (ripetizioni sospette), 2 (ripetizioni evidenti)
<i>Qualità</i>	Misura la correttezza complessiva del testo, evidenziando omissioni o formulazioni improprie	Da 1 (qualità scarsa, con incongruenze o omissioni che compromettono la chiarezza) a 4 (qualità ottima)
<i>Struttura</i>	Valuta la presenza e completezza delle sezioni obbligatorie	SI (struttura adeguata) oppure NO (non adeguata)

3.5. Processo iterativo di confronto e affinamento

I referti forniti in input al sistema sono documenti clinici reali prodotti presso il Centro Diagnostico Italiano. Prima dell'analisi automatica, i documenti vengono convertiti in formato testuale e sottoposti a procedure di anonimizzazione, al fine di rimuovere o mascherare le informazioni identificative del paziente, in conformità con i requisiti di protezione dei dati personali. A ciascun referto viene quindi associato un identificativo, necessario per consentire il successivo confronto tra output algoritmico e valutazione manuale.



Figura 2: Pre-processing dei referti clinici.

L'algoritmo legge e analizza i referti, restituendo in output una valutazione strutturata dei parametri richiesti, organizzata in forma tabellare. Tali output possono essere esportati e successivamente elaborati per l'analisi quantitativa. Prima del confronto con la valutazione umana, i risultati prodotti dal sistema sono sottoposti a una fase di post-processing da parte del team di ReportAId, finalizzata a organizzare, normalizzare e rendere confrontabili gli output rispetto ai criteri di valutazione definiti. I dati così ottenuti costituiscono la base informativa per il calcolo dei KPI e per l'analisi dell'allineamento tra valutazione automatica e giudizio umano.

Il processo adottato da ReportAId si basa su un meccanismo iterativo di confronto tra valutazione automatica e valutazione manuale. Nelle prime fasi del progetto, il CDI ha trasmesso progressivamente gruppi di referti già esaminati da un valutatore umano esperto. Tali referti sono stati analizzati dall'algoritmo e i risultati ottenuti sono stati confrontati con il giudizio manuale, al fine di individuare eventuali discrepanze e fornire feedback utili al miglioramento del sistema.

I feedback derivanti da questo confronto sono stati utilizzati per affinare progressivamente il modello, aumentandone l'allineamento rispetto ai criteri di valutazione richiesti. La procedura è stata ripetuta su dataset successivi, aggiornati in funzione delle nuove valutazioni manuali disponibili, permettendo al sistema di incorporare progressivamente le indicazioni emerse dalla revisione umana.

A partire dal quarto dataset utilizzato, è stato adottato un approccio metodologico differente. Dopo le precedenti fasi di affinamento, il modello è stato applicato a un nuovo set di referti non ancora valutati manualmente.

Gli output prodotti dall'algoritmo sono stati quindi inviati alla Direttrice Sanitaria, che ha effettuato la revisione manuale degli stessi referti avendo accesso alle valutazioni e alle motivazioni fornite dall'AI. Anche in questa fase, il confronto tra giudizio umano e output algoritmico ha permesso di effettuare un ulteriore affinamento del modello, sulla base del bilancio tra valutazione manuale e valutazione automatica.



Figura 3: Schema di processo iterativo di affinamento del modello.

3.6. Dataset disponibili e dataset inclusi nell'analisi quantitativa

Nel corso del progetto sono stati prodotti diversi report successivi, riferiti a campioni di referti appartenenti a periodi temporali differenti. Tali report documentano l'evoluzione del sistema ReportAId e permettono di osservare l'andamento della concordanza tra valutazione automatica e valutazione umana sui principali parametri analizzati.

Tabella 5: Elementi principali dei sei report.

Report	Data report	Periodo dei referti	Numero di referti analizzati	Finalità principale
Report 1	23/07/2025	Marzo 2025	125	Prima valutazione delle prestazioni di ReportAId
Report 2	02/09/2025	Aprile–giugno 2025	708	Monitoraggio delle prestazioni su campione più ampio

Report 3	24/09/2025	Aprile–giugno 2025	708	Aggiornamento metodologico e rivalutazione degli acronimi
Report 4	28/01/2026	Luglio–settembre 2025	89	Analisi preliminare con nuovo approccio metodologico
Report 5	26/03/2026	Luglio–settembre 2025	370	Validazione indipendente su dati non osservati
Report 6	09/04/2026	Luglio–settembre 2025	161	Ulteriore validazione

Ai fini dell’analisi quantitativa dei KPI, tuttavia, sono stati considerati esclusivamente i dataset per i quali era disponibile una corrispondenza certa tra output AI e valutazione manuale mediante identificativo del referto. Questa scelta ha permesso di basare il confronto su dati allineati e metodologicamente più solidi.

Sono stati quindi analizzati due insiemi di dati:

- un primo dataset preliminare, costituito da 89 referti, utilizzato anche nella fase di affinamento del modello;
- un dataset finale di validazione, costituito da 161 referti, assunto come riferimento principale per la valutazione delle prestazioni della soluzione AI.

Nel primo dataset, pur essendo presenti 89 righe complessive, gli identificativi univoci risultano pari a 87 a causa di due duplicazioni del campo report_ID. Tali righe sono state mantenute nell’elaborazione per preservare la coerenza con la numerosità riportata nella documentazione progettuale. Il dataset finale di validazione presenta invece 161 righe e 161 identificativi univoci.

I due dataset sono stati analizzati separatamente e non aggregati, poiché riferiti a fasi differenti del progetto e caratterizzati da ruoli metodologici distinti. Il primo dataset è

stato utilizzato come riferimento descrittivo preliminare, mentre il dataset finale rappresenta la base principale per la valutazione conclusiva del potenziale supporto organizzativo della soluzione AI.

3.7. Pre-processing e matching AI–umano

Prima del calcolo dei KPI, i dati sono stati sottoposti a una fase di pre-processing. Per ciascun dataset sono state considerate due tabelle: una contenente le valutazioni prodotte dal sistema AI e una contenente le corrispondenti valutazioni manuali.

Le due tabelle, come scritto precedentemente, sono state allineate mediante identificativo del referto, in modo da associare ogni output algoritmico al giudizio umano riferito allo stesso documento. I casi privi di una corrispondenza verificabile sono stati esclusi dall'analisi quantitativa.

Successivamente, le variabili sono state convertite in formato numerico per consentire il calcolo degli indicatori. La qualità è stata mantenuta come variabile ordinale su scala da 1 a 4; il Copia-Incolla improprio è stato trattato come variabile ordinale su scala da 0 a 2; acronimi e struttura sono stati convertiti in variabili binarie.

3.8. Normalizzazione delle variabili

Poiché i quattro parametri analizzati presentano scale di misura differenti, è stata effettuata una normalizzazione delle variabili su scala comune compresa tra 0 e 1. La normalizzazione è stata orientata in modo tale che il valore 1 rappresenti sempre la condizione migliore e il valore 0 la condizione peggiore. La Qualità è stata normalizzata secondo la formula:

$$Q_n = \frac{Q - 1}{3}$$

dove Q rappresenta il punteggio originario di qualità, compreso tra 1 e 4.

Per la variabile relativa agli Acronimi è stata calcolata una misura positiva corrispondente all'assenza di errore:

$$A_n = 1 - A$$

dove $A = 1$ indica la presenza di acronimi o simboli non consentiti e $A = 0$ indica la loro assenza.

Per il Copia-Incolla, essendo il valore 0 associato alla condizione migliore e il valore 2 alla condizione peggiore, è stata utilizzata la formula:

$$C_n = 1 - \frac{C}{2}$$

Infine, per la Struttura è stato mantenuto direttamente il valore binario normalizzato:

$$S_n = S$$

dove $S = 1$ indica una struttura corretta e $S = 0$ una struttura non corretta.

3.9. Definizione dei KPI

Per valutare in modo quantitativo gli output prodotti dal sistema AI è stato definito un insieme di Key Performance Indicators (KPI). Gli indicatori sono stati suddivisi in categorie principali:

- KPI descrittivi dell'output AI;
- KPI avanzati di qualità;
- KPI di concordanza tra valutazione AI e valutazione umana;
- KPI organizzativi per l'identificazione dei referti da revisionare.

I KPI descrittivi includono la **qualità media** dei referti erogati, la percentuale di referti con **qualità massima**, la percentuale di referti con **acronimi o simboli non consentiti**, la percentuale di referti con **struttura corretta** e il valore medio di **copy-paste improprio**.

Tabella 6: Kpi descrittivi di base.

<i>KPI</i>	<i>DEFINIZIONE</i>
<i>Mean_Q</i>	Qualità media dei referti
<i>Perc_Q4</i>	Percentuale di referti con qualità massima pari a 4
<i>Acr_Err_%</i>	Percentuale di referti con acronimi o simboli non consentiti
<i>Struttura_OK_%</i>	Percentuale di referti con struttura corretta
<i>Mean_Copy</i>	Valore medio del copia-incolla improprio

Per sintetizzare la qualità complessiva del referto è stato definito un **Global Quality Score**, calcolato come somma pesata delle quattro variabili normalizzate:

$$GQS = 0.40Q_n + 0.15A_n + 0.15C_n + 0.30S_n$$

I pesi sono stati assegnati attribuendo maggiore importanza alla qualità complessiva del referto e alla correttezza della struttura, considerate dimensioni centrali nella valutazione della qualità documentale. Acronimi e Copia-Incolla sono stati inclusi con peso inferiore, ma comunque rilevante, in quanto rappresentano elementi specifici di inappropriatezza o ridondanza del referto.

Tra i KPI avanzati sono stati inclusi l'**Error Burden Index**, la **Compliance**, la percentuale di referti **Almost Perfect**, il **Quality Consistency Index** e il **Consistency Ratio**.

Tabella 7: KPI avanzati.

<i>KPI</i>	<i>DEFINIZIONE</i>
<i>Error Burden Index</i>	Carico medio di errore legato alla presenza di acronimi o simboli non consentiti e copia-incolla improprio
<i>Compliance</i>	Percentuale di referti pienamente conformi ($Q=4, A=0, C=0, S=1$)
<i>Almost perfect</i>	Percentuale di referti complessivamente buoni anche in presenza di criticità minori ($Q \geq 3, A = 0, C \leq 1, S = 1$)
<i>Quality Consistency Index</i>	Valuta la stabilità della qualità degli output AI
<i>Consistency Ratio</i>	Rapporto tra la media del GQS e la sua deviazione standard

L'**Error Burden Index** misura il carico medio di errore legato alla presenza di acronimi o simboli non consentiti e copy-paste improprio:

$$EBI = mean\left(A + \frac{C}{2}\right)$$

La **Compliance** identifica la percentuale di referti pienamente conformi, ossia referti che soddisfano simultaneamente le condizioni:

$$Q = 4, A = 0, C = 0, S = 1$$

È stato inoltre definito un criterio meno restrittivo, denominato **Almost Perfect**, che identifica i referti complessivamente buoni anche in presenza di criticità minori:

$$Q \geq 3, A = 0, C \leq 1, S = 1$$

Il **Quality Consistency Index** è stato utilizzato per valutare la stabilità della qualità degli output AI:

$$QCI = 1 - std(GQS)$$

Il **Consistency Ratio** è stato invece calcolato come rapporto tra la media del GQS e la sua deviazione standard. Un valore elevato di tale indice indica che il modello assegna punteggi mediamente alti e, al contempo, caratterizzati da una ridotta variabilità, suggerendo quindi una maggiore coerenza nelle valutazioni prodotte. Al contrario, un valore basso può riflettere punteggi medi più contenuti e/o una maggiore dispersione dei risultati, evidenziando una minore stabilità del processo valutativo.

$$CR = \frac{mean(GQS)}{std(GQS)}$$

3.10. Metriche di concordanza tra AI e valutazione umana

Per valutare il grado di allineamento tra la valutazione automatica e il giudizio umano, sono state calcolate metriche di concordanza specifiche per ciascun parametro.

Tabella 8: KPI di concordanza tra AI e valutazione umana.

<i>KPI</i>	<i>DEFINIZIONE</i>
<i>Mean Absolute Error</i>	Scostamento medio assoluto tra la valutazione automatica e quella manuale
<i>Concordanza Tollerante</i>	Scostamento massimo di un livello accettabile tra la valutazione automatica e quella manuale
<i>Full Match</i>	Percentuale di referti per cui AI e umano coincidono simultaneamente su tutti e quattro i parametri
<i>Critical Discordance</i>	Percentuale di casi in cui la discrepanza tra AI e umano può essere considerata rilevante
<i>Cohen's Kappa</i>	Accordo tra AI e umano su variabili binarie, corretto per l'accordo casuale.
<i>Weighted Kappa</i>	Accordo tra AI e umano su variabili ordinali, considerando la gravità dello scostamento.

Per le variabili ordinali, qualità e copia-incolla, è stato calcolato il **Mean Absolute Error**, che misura lo scostamento medio assoluto tra la valutazione automatica e quella manuale. Viene definito come:

$$MAE = mean(| AI - Umano |)$$

Per le stesse variabili è stata inoltre calcolata una **concordanza tollerante**, considerando accettabile uno scostamento massimo di un livello:

$$Accuracy_{tol1} = mean(| AI - Umano | \leq 1) \times 100$$

Per valutare la completa sovrapposibilità del profilo valutativo attribuito al singolo referto, è stato calcolato il **Full Match**, definito come la percentuale di referti per cui AI e umano coincidono simultaneamente su tutti e quattro i parametri.

$$Full Match = mean(AI_A = H_A \cap AI_S = H_S \cap AI_Q = H_Q \cap AI_C = H_C) \times 100$$

È stata inoltre definita una **Critical Discordance**, cioè la percentuale di casi in cui la discrepanza tra AI e umano può essere considerata rilevante dal punto di vista organizzativo. Sono state considerate critiche le situazioni in cui si verifica una discordanza sulla presenza di acronimi, una discordanza sulla struttura oppure una differenza di almeno due livelli nel punteggio di qualità.

Per le variabili binarie, acronimi e struttura, è stato calcolato il **Cohen's Kappa**, che misura l'accordo tra due valutatori correggendo per l'accordo atteso casualmente. Il Cohen's kappa è più informativo della semplice accuratezza, poiché tiene conto della possibilità che una parte dell'accordo tra due valutatori avvenga casualmente.

L'indice è definito come:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

dove p_o rappresenta la proporzione di accordo osservato tra valutazione automatica e valutazione umana, mentre p_e rappresenta la proporzione di accordo attesa per effetto del caso. In particolare:

$$p_o = \sum_{i=1}^K p_{ii}$$

$$p_e = \sum_{i=1}^K p_{i.} \cdot p_{.i}$$

dove K indica il numero di categorie considerate, p_{ii} rappresenta la proporzione di casi per cui entrambi i valutatori assegnano la categoria i , mentre $p_{i.}$ e $p_{.i}$ rappresentano le proporzioni marginali attribuite rispettivamente dall'AI e dal valutatore umano alla medesima categoria.

Nel presente studio, essendo Acronimi e Struttura variabili binarie, il numero di categorie considerate è pari a $K = 2$.

Per le variabili ordinali è stato invece calcolato il **Weighted Kappa** lineare, più adatto a variabili su scala ordinata perché tiene conto della distanza tra le categorie. In questo modo, una discrepanza di un solo livello viene penalizzata meno rispetto a una discrepanza di due o più livelli.

Il Weighted Kappa è definito come:

$$\kappa_w = \frac{p_o^{(w)} - p_e^{(w)}}{1 - p_e^{(w)}}$$

dove $p_o^{(w)}$ rappresenta l'accordo osservato ponderato e $p_e^{(w)}$ l'accordo atteso ponderato. Tali quantità sono calcolate come:

$$p_o^{(w)} = \sum_{i=1}^k \sum_{j=1}^K w_{ij} p_{ij}$$

$$p_e^{(w)} = \sum_{i=1}^K \sum_{j=1}^K w_{ij} p_{i.p.j}$$

Nel caso della ponderazione lineare, i pesi sono definiti secondo la relazione:

$$w_{ij} = 1 - \frac{(|i - j|)}{(K - 1)}$$

dove i e j rappresentano rispettivamente la categoria attribuita dall'AI e quella attribuita dal valutatore umano, mentre K indica il numero complessivo di livelli della variabile. Il peso assume valore massimo in caso di perfetta concordanza e si riduce progressivamente all'aumentare della distanza tra le categorie assegnate.

Nel presente studio, il Weighted Kappa lineare è stato applicato al parametro Qualità, articolato su quattro livelli ($K = 4$), e al parametro Copia-Incolla, articolato su tre livelli ($K = 3$).

3.11. KPI organizzativi per l'identificazione dei referti da revisionare

Oltre alla valutazione della concordanza sui singoli parametri, è stata definita una classificazione organizzativa dei referti come “**da revisionare**” o “**non da revisionare**”. Un referto è stato considerato da revisionare nel caso in cui presentasse almeno una non piena conformità rispetto ai criteri analizzati, ovvero qualità inferiore al punteggio massimo, presenza di acronimi o simboli non consentiti, copy-paste improprio maggiore di zero oppure struttura non corretta.

Questa classificazione non identifica necessariamente referti clinicamente inadeguati, ma individua documenti che, secondo il sistema di controllo qualità adottato, potrebbero richiedere un'ulteriore verifica.

Sulla base di questa classificazione sono stati calcolati indicatori tipici di un sistema di triage, tra cui **precision**, **recall**, **accuracy**, **specificità**, **F1-score**, **falsi positivi** e **falsi negativi**, **veri positivi** e **veri negativi**. In questo contesto, il recall assume particolare rilevanza, poiché misura la capacità del sistema di intercettare i referti che il valutatore umano considera meritevoli di revisione. Una sensibilità elevata è infatti desiderabile in un sistema di pre-screening, in quanto riduce il rischio di non segnalare documenti potenzialmente problematici.

Tabella 9: KPI organizzativi.

<i>KPI</i>	<i>SIGNIFICATO</i>
<i>True positive</i>	Referti critici sia secondo AI sia secondo umano
<i>True negative</i>	Referti non critici sia secondo AI sia secondo umano
<i>False positive</i>	Referti segnalati come critici dall'AI ma non dal valutatore umano
<i>False negative</i>	Referti non segnalati come critici dall'AI ma critici secondo il valutatore umano
<i>Precision</i>	Affidabilità delle segnalazioni AI
<i>Recall</i>	Capacità dell'AI di intercettare i referti critici
<i>Specificità</i>	Capacità dell'AI di riconoscere i referti non critici
<i>F1-score</i>	Indicatore sintetico tra precision e recall

3.12. Implementazione dell'analisi in MATLAB

L'elaborazione dei dati è stata effettuata mediante uno script sviluppato in MATLAB, utilizzato per rendere il processo di analisi quantitativa automatizzato e riproducibile. Il codice ha previsto l'importazione dei dataset contenenti le valutazioni prodotte dall'AI e quelle effettuate manualmente, la selezione dei referti per i quali era disponibile una corrispondenza certa tra output algoritmico e valutazione umana, la conversione delle variabili in formato numerico e la normalizzazione dei parametri necessari al calcolo degli indicatori.

A partire dai dati rielaborati, lo script ha consentito di calcolare gli indicatori descrittivi e le principali metriche di confronto tra valutazione automatica e valutazione umana. In particolare, sono stati ottenuti KPI relativi alla qualità globale dei referti, alla presenza di criticità documentali, alla compliance rispetto ai criteri definiti, alla concordanza AI–umano e alla capacità del sistema di individuare i referti meritevoli di revisione.

L'utilizzo di MATLAB ha permesso di ridurre il rischio di errori manuali nelle fasi di calcolo e di garantire una maggiore coerenza metodologica tra i dataset analizzati. I risultati ottenuti sono stati successivamente organizzati in tabelle e grafici riassuntivi, utilizzati per l'interpretazione dei dati e per la discussione dei risultati.

Nel presente capitolo vengono descritte le principali scelte metodologiche adottate per la costruzione e l'interpretazione dei KP

Capitolo 4

Risultati

4.1. Report progettuali e andamento generale della concordanza AI–umano

Il confronto tra la valutazione automatica e la valutazione umana è stato effettuato sui quattro parametri principali considerati nell'analisi: Acronimi, Copia-Incolla, Struttura e Qualità. Per i parametri binari o categoriali (Acronimi e Struttura) la concordanza è stata calcolata come percentuale di casi in cui il giudizio dell'algoritmo coincideva con quello umano. Per il parametro Qualità, espresso su una scala da 1 a 4, sono state considerate sia la concordanza esatta sia la concordanza entro uno scostamento massimo di un punto. Quest'ultima misura consente di tenere conto della natura parzialmente soggettiva del parametro, distinguendo le divergenze rilevanti da quelle riconducibili a lievi differenze interpretative.

Nel complesso, i risultati dei report rappresentati nella *Tabella 10* mostrano un andamento non perfettamente lineare, ma indicativo di un progressivo affinamento del sistema. Tale andamento evidenzia come l'ottimizzazione del modello non dipenda esclusivamente dalle sue capacità tecniche, ma anche dalla chiarezza delle istruzioni fornite e dalla definizione condivisa dei criteri di valutazione. Alcuni parametri, come il Copia-Incolla, presentano prestazioni elevate e stabili già dalle prime fasi, probabilmente in quanto basati su caratteristiche più facilmente identificabili in modo oggettivo. Altri, come Acronimi e Struttura, risultano maggiormente sensibili alla definizione delle regole e alla modalità di applicazione dei criteri da parte del valutatore umano. Il parametro Qualità si

conferma, invece, il più complesso, poiché implica una valutazione più soggettiva e multidimensionale del contenuto del referto. La valutazione di tale dimensione richiede infatti di integrare aspetti di completezza, chiarezza espositiva, appropriatezza informativa e utilità clinica, rendendo più probabile uno scostamento tra giudizio umano e giudizio automatizzato. I sottocapitoli successivi presentano l'analisi degli andamenti delle percentuali tra la valutazione umana e quella automatica, parametro per parametro.

Tabella 10: Percentuali di concordanza tra valutazione automatica e umana.

<i>Report</i>	<i>N. Referti</i>	<i>Acronimi</i>	<i>Copy & Paste</i>	<i>Struttura</i>	<i>Qualità Esatta</i>	<i>Qualità 1</i>
23/07/2025	125	28,0%	96,0%	67,2%	28,8%	77,6%
02/09/2025	708	35,7%	97,7%	56,8%	57,2%	97,0%
24/09/2025	708	38,4%*	97,7%	56,8%	57,2%	97,0%
28/01/2026	89	87,64%	86,52%	91,01%	53,93%	100%
26/03/2026	370	74,59%	97,84%	82,70%	65,41%	97,03%
09/04/2026	161	69,57%	89,06%	81,37%	55,90%	68,3%

4.1.1. Analisi del parametro Acronimi

Il parametro Acronimi valuta la presenza di sigle, abbreviazioni o simboli non consentiti secondo le linee guida interne del CDI. Nei primi report, questo parametro ha mostrato livelli di concordanza relativamente bassi tra valutazione automatica e valutazione umana. Tale risultato non indica necessariamente una scarsa capacità del sistema, ma riflette soprattutto una diversa severità tra algoritmo e revisore umano.

L'algoritmo tende, infatti, a segnalare in modo sistematico tutte le occorrenze potenzialmente non conformi, incluse sigle clinicamente comprensibili o comunemente tollerate dal personale sanitario, ma non formalmente presenti nella lista degli acronimi autorizzati. Questo comportamento determina un aumento delle segnalazioni prodotte

dall'AI e, di conseguenza, una possibile riduzione della concordanza con il giudizio umano. Nel corso del progetto sono state introdotte revisioni metodologiche, tra cui l'ampliamento della lista degli acronimi consentiti e una rivalutazione dell'impatto delle sigle di laboratorio. Le analisi successive hanno mostrato un netto miglioramento della concordanza rispetto alle fasi iniziali, pur mantenendo alcune aree di disaccordo.

Nel complesso, il parametro Acronimi evidenzia un miglioramento rilevante nel tempo, ma rimane fortemente dipendente dalla definizione condivisa di quali abbreviazioni debbano essere considerate accettabili nel contesto clinico CDI. Il principale margine di ottimizzazione riguarda quindi l'aggiornamento progressivo della lista degli acronimi consentiti, integrando le sigle emerse come clinicamente accettabili durante la revisione congiunta con gli esperti.

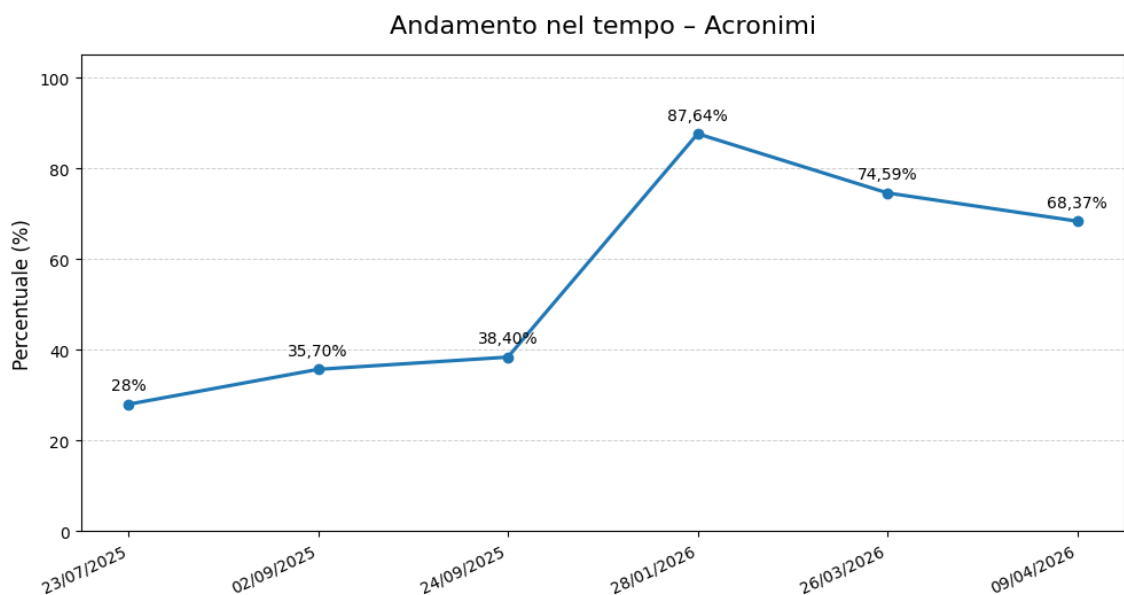


Figura 4: Andamento nel tempo della concordanza AI-Umano del parametro Acronimi.

4.1.2. Analisi del parametro Copia-Incolla

Il parametro Copia-Incolla valuta la presenza di contenuti duplicati, ripetizioni anomale o elementi fuori contesto che possano suggerire un uso improprio del Copia-Incolla nella redazione del referto.

Questo parametro presenta, nella maggior parte dei report, le prestazioni più elevate e stabili. La concordanza tra AI e valutazione umana risulta generalmente alta, suggerendo che il sistema sia in grado di riconoscere correttamente l'assenza di anomalie nella maggior parte dei referti analizzati.

Le principali discordanze osservate sono riconducibili soprattutto a falsi positivi, cioè casi in cui l'algoritmo segnala come sospette ripetizioni o formulazioni standard che il valutatore umano considera invece accettabili nel contesto clinico. Queste situazioni possono riguardare frasi ricorrenti previste dalla struttura del referto, formulazioni standardizzate o modalità redazionali specifiche di alcune specialità.

Un aspetto importante riguarda la bassa frequenza di veri casi di Copia-Incolla improprio nei dataset analizzati. Questo elemento rappresenta un limite nella valutazione del parametro, poiché riduce la possibilità di testare in modo completo la capacità discriminativa del sistema. Pertanto, le prestazioni elevate devono essere interpretate tenendo conto della distribuzione dei casi nel campione. Nel complesso, Copia-Incolla risulta il parametro più stabile del sistema, ma ulteriori verifiche su dataset contenenti un maggior numero di casi realmente patologici sarebbero utili per valutarne in modo più robusto la capacità di identificare anomalie rilevanti.

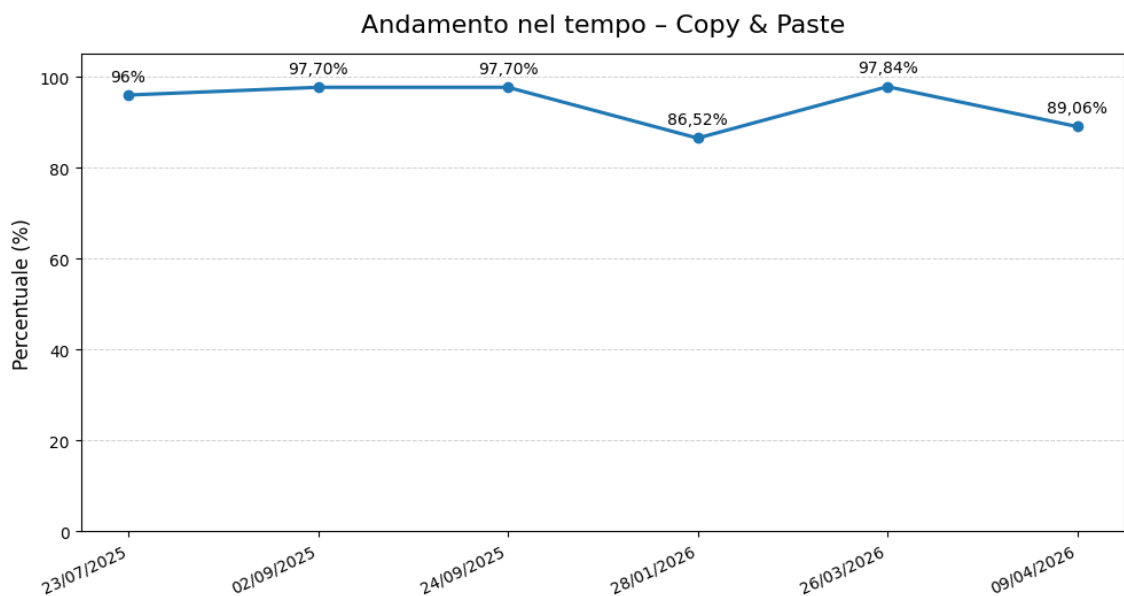


Figura 5: Andamento nel tempo della concordanza AI-Umano del parametro Copia-Incolla.

4.1.3. Analisi del parametro Struttura

Il parametro Struttura valuta l'organizzazione redazionale del referto, con particolare attenzione alla presenza delle sezioni obbligatorie, alla loro corretta compilazione e alla collocazione appropriata delle informazioni cliniche. Nei primi report, la concordanza tra AI e valutazione umana risulta più variabile rispetto al parametro Copia-Incolla. Questa variabilità è riconducibile alla maggiore complessità del criterio, che richiede di distinguere tra errori strutturali effettivi e prassi redazionali specifiche di alcune discipline cliniche.

In alcune situazioni, il sistema tende ad applicare i criteri in modo più rigoroso rispetto al valutatore umano, classificando come non conformi referti che presentano deviazioni formali, ma che possono essere considerate accettabili nella pratica clinica. In altri casi, invece, il sistema può non rilevare errori strutturali considerati rilevanti dal revisore umano, come la presenza di informazioni terapeutiche o di follow-up collocate in sezioni non appropriate. Nelle fasi più recenti del progetto, la concordanza sul parametro Struttura mostra un miglioramento rispetto alle prime analisi. Tuttavia, le criticità residue suggeriscono la necessità di integrare nel modello alcune specificità redazionali proprie delle diverse specialità cliniche, in modo da distinguere meglio tra errore strutturale effettivo e variabilità accettabile nella compilazione del referto.

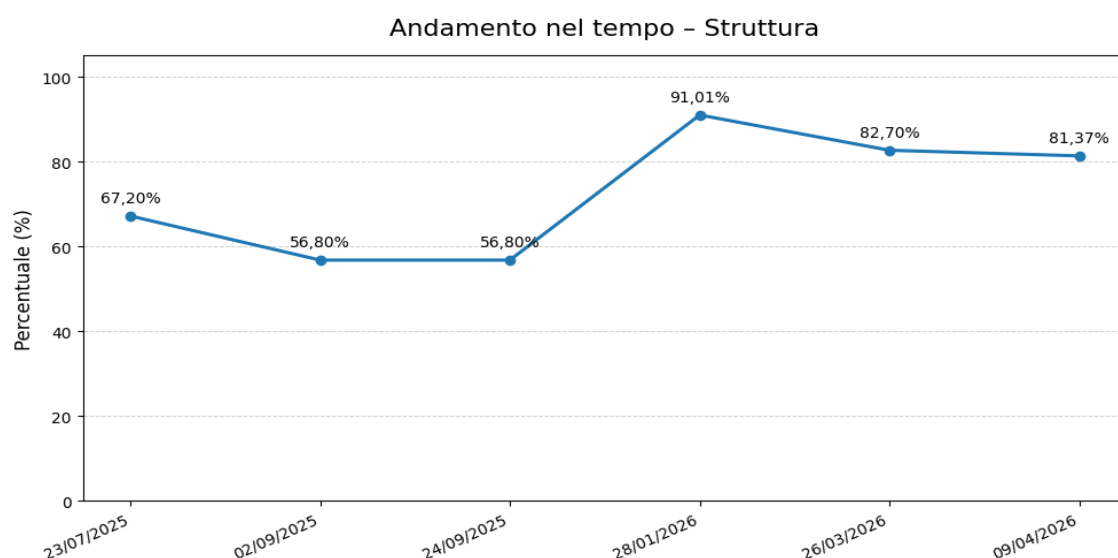


Figura 6: Andamento nel tempo della concordanza AI-Umano del parametro Struttura.

4.1.4. Analisi del parametro Qualità

Il parametro Qualità rappresenta l'indicatore più complesso tra quelli analizzati, poiché richiede una valutazione complessiva della correttezza redazionale del referto. A differenza degli altri parametri, non si basa esclusivamente su elementi binari o facilmente oggettivabili, ma implica l'interpretazione di aspetti quali chiarezza, completezza, coerenza, appropriatezza delle informazioni e livello di dettaglio clinico.

La concordanza esatta tra valutazione AI e valutazione umana risulta generalmente più bassa e variabile rispetto agli altri parametri. Tuttavia, quando si considera la concordanza entro uno scostamento massimo di un punto, i valori risultano sensibilmente più elevati. Questo suggerisce che, anche quando il sistema non assegna esattamente lo stesso punteggio del valutatore umano, nella maggior parte dei casi si colloca comunque vicino al giudizio esperto.

Le discrepanze osservate sul parametro Qualità sembrano quindi dipendere soprattutto da differenze di severità valutativa, più che da errori completamente discordanti. Tale risultato è coerente con la natura ordinale e soggettiva del parametro, che richiede una valutazione più interpretativa rispetto a criteri come Acronimi o Copia-Incolla. Nel complesso, la Qualità rimane l'ambito più complesso da automatizzare.

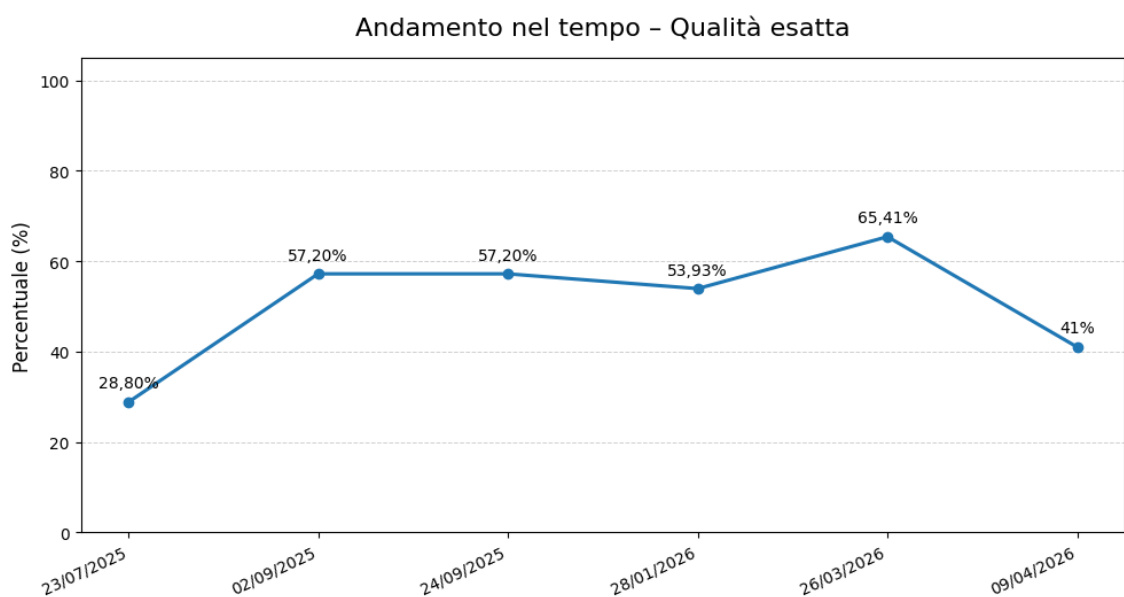


Figura 7: Andamento nel tempo della concordanza AI-Umano del parametro Qualità esatta.

4.2. Analisi dei KPI calcolati

Avendo a disposizione, per entrambi i dataset analizzati, sia la valutazione umana sia quella prodotta dall'algoritmo sui quattro parametri considerati — qualità, struttura, acronimi non consentiti e copia-incolla improprio — è stato possibile calcolare diversi indicatori KPI, utilizzati con due obiettivi: descrivere e confrontare gli output dell'AI nelle due fasi analizzate e valutare il grado di concordanza tra algoritmo e giudizio umano.

4.2.1. KPI descrittivi degli output AI

La prima analisi quantitativa ha riguardato la caratterizzazione delle valutazioni prodotte dall'algoritmo. Questa fase ha avuto lo scopo di descrivere il comportamento del sistema di valutazione automatica e di fornire al team qualità una prima rappresentazione sintetica delle principali criticità rilevate nei referti clinici. In particolare, tale analisi risulta utile per individuare gli aspetti della refertazione sui quali i medici del CDI possono eventualmente intervenire con maggiore attenzione, sia in termini di qualità complessiva del documento sia in relazione ad aspetti più specifici, come l'utilizzo di acronimi, la corretta organizzazione della struttura e la presenza di copy-paste improprio.

Nel set finale di validazione da 161 referti, la **qualità media** assegnata dall'AI è risultata pari a **3,52** su una scala da 1 a 4, con una **mediana** pari a **4**. Inoltre, il **55,90%** dei referti è stato valutato con il punteggio massimo di **qualità**. Questo dato evidenzia che, secondo l'algoritmo, oltre la metà dei referti appartenenti all'ultimo set di validazione presenta una qualità complessiva molto elevata. La presenza di una mediana pari al valore massimo conferma ulteriormente che la distribuzione delle valutazioni tende a concentrarsi verso i punteggi più alti, suggerendo un quadro complessivamente positivo per quanto riguarda la qualità generale dei documenti analizzati.

Per quanto riguarda le criticità redazionali, l'AI ha rilevato la presenza di **acronimi o simboli non consentiti** nel **69,57%** dei referti. Questo dato indica che l'uso di abbreviazioni, sigle o simboli non conformi rappresenta ancora una delle principali aree di attenzione nella stesura dei referti clinici.

La percentuale di referti giudicati **strutturalmente corretti** è risultata pari al **46,58%**, suggerendo che la struttura del referto rappresenta ancora un ambito caratterizzato da un

marginale di miglioramento. Tale aspetto può riguardare, ad esempio, l'organizzazione delle informazioni, la completezza delle sezioni previste o la coerenza formale del documento rispetto agli standard attesi.

Il **copia-incolla improprio**, invece, presenta un valore medio contenuto, pari a **0,07** su una scala da 0 a 2, indicando che tale criticità risulta quantitativamente meno rilevante rispetto ad acronimi e struttura.

Nel dataset preliminare, la **qualità media** assegnata dall'AI era pari a **3,06**, con una percentuale di referti valutati con **qualità massima** pari al **7,87%**. L'algoritmo segnalava **acronimi o simboli non consentiti** nell'**82,02%** dei casi, mentre solo il **17,98%** dei referti risultava strutturalmente corretto. Anche il livello medio di **copia-incolla** risultava più elevato, pari a **0,20**. Nel complesso, i dati descrittivi del set finale mostrano un profilo qualitativo più favorevole rispetto alla fase preliminare. Tuttavia, tale confronto deve essere interpretato in termini descrittivi e non come dimostrazione causale di un miglioramento del modello, poiché i due dataset hanno ruolo e composizione differenti.

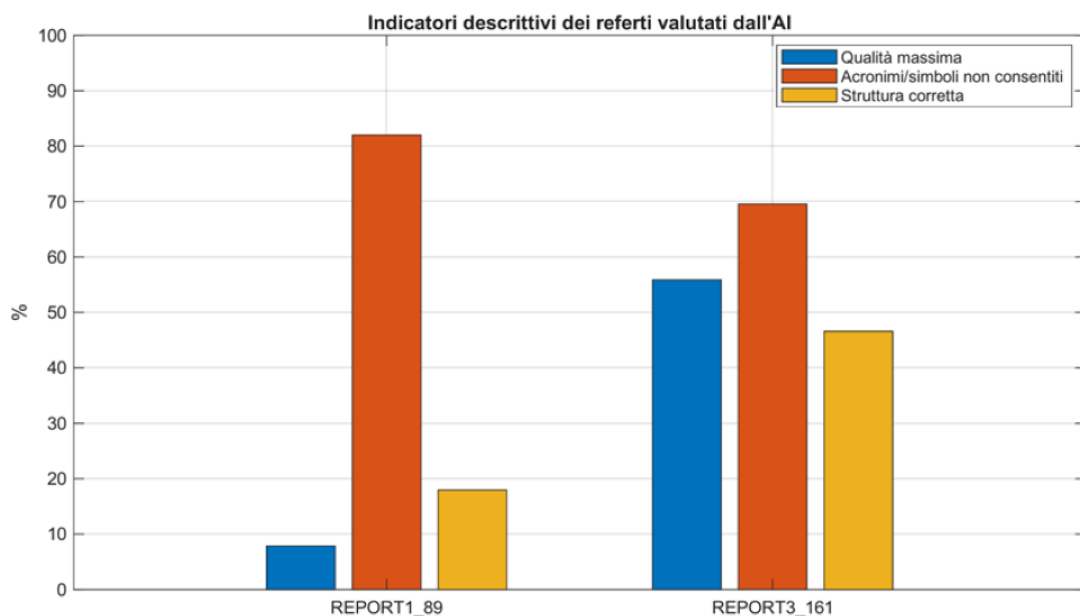


Figura 8: Confronto degli indicatori percentuali rilevati dall'algoritmo nei dataset preliminare e finale di validazione.

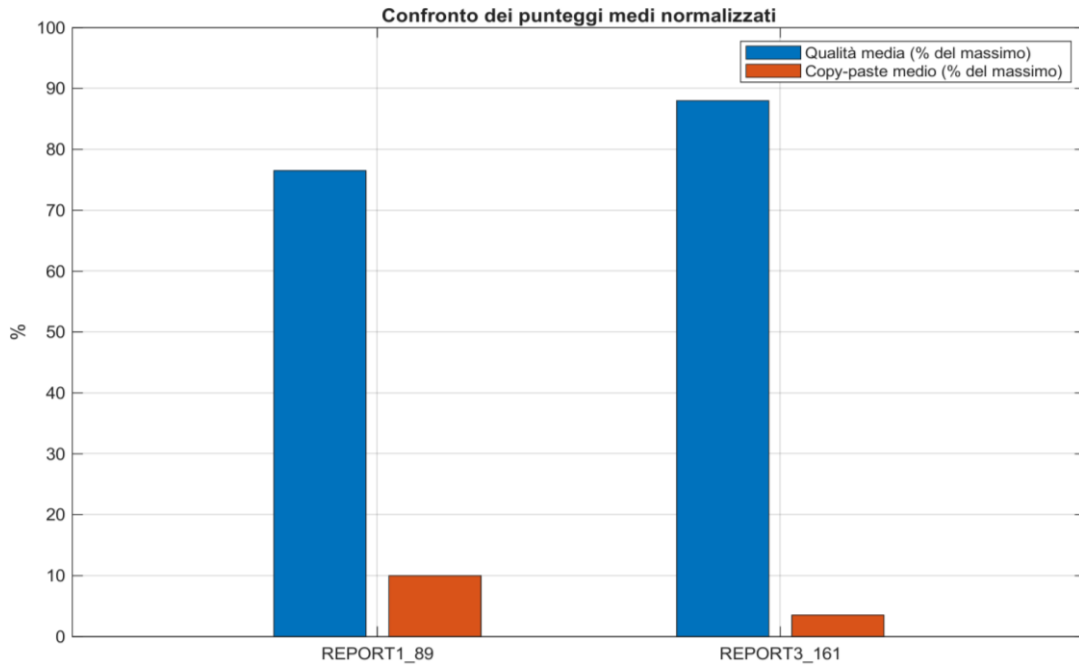


Figura 9: Confronto dei punteggi medi normalizzati relativi alla qualità dei referti e al livello di copia-incolla improprio nei due dataset analizzati.

4.2.2. KPI avanzati di qualità dell'output AI

Per ottenere una lettura sintetica e multidimensionale della qualità dei referti valutati dall'algoritmo, sono stati calcolati alcuni indicatori avanzati, finalizzati a integrare le informazioni fornite dai singoli parametri analizzati. Oltre alla valutazione puntuale di qualità, acronimi, copia-incolla e struttura, questi indicatori consentono infatti di rappresentare in modo più complessivo il profilo qualitativo dei referti e di sintetizzare, attraverso metriche aggregate, il livello di conformità e di stabilità delle valutazioni prodotte dall'AI.

In particolare, sono stati considerati l'**Error Burden Index**, **Compliance**, **Almost Perfect**, **Global Quality Score**, **Quality Consistency Index** e **Consistency Ratio**. Ciascuno di questi indici descrive un aspetto differente della qualità del referto: alcuni permettono di quantificare il carico complessivo di criticità, altri misurano la quota di referti pienamente o quasi pienamente conformi, mentre altri ancora descrivono il livello medio di qualità globale e la sua omogeneità all'interno del dataset.

Nel set finale di validazione, l'**Error Burden Index** è risultato pari a **0,730**. Questo valore indica la presenza di un carico medio di criticità non trascurabile, legato principalmente alla rilevazione di acronimi o simboli non consentiti e, in misura minore, al Copia-Incolla improprio. L'EBI permette quindi di evidenziare come, pur in presenza di una qualità media complessivamente buona, permangano alcune criticità redazionali che incidono sul livello di conformità complessivo dei documenti.

La percentuale di referti classificati come pienamente conformi secondo la definizione più stringente, espressa dalla **Compliance**, è risultata pari al **6,21%**. Ciò significa che una quota limitata di referti soddisfa contemporaneamente tutte le condizioni di piena conformità considerate: qualità massima, assenza di acronimi o simboli non consentiti, assenza di copy-paste improprio e struttura corretta.

Applicando un criterio meno restrittivo, la percentuale di referti **Almost Perfect** raggiunge il **15,53%**. Questo indicatore evidenzia che, pur rimanendo ridotta la quota di referti perfettamente conformi, una percentuale più ampia si colloca in una fascia di qualità complessivamente buona, con criticità minori o contenute.

Il **Global Quality Score medio** del set finale di validazione è risultato pari a **0,667** su scala 0-1. Tale valore suggerisce una qualità globale complessivamente discreta secondo la valutazione AI, derivante dalla combinazione ponderata dei quattro criteri analizzati.

La deviazione standard del **GQS** è pari a **0,175**, mentre il **Quality Consistency Index** risulta pari a **0,825**, indicando una discreta omogeneità nella distribuzione della qualità stimata dall'algoritmo. Il **Consistency Ratio**, pari a **3,80**, conferma una buona combinazione tra livello medio del punteggio e stabilità della distribuzione.

Nel Report 1, l'**EBI** era pari a **0,921**, la **Compliance** era **nulla** e la quota di referti **quasi perfetti** era pari all'**1,12%**. Il **GQS** medio risultava pari a **0,490**, con **QCI** pari a **0,868** e **CR** pari a **3,70**. Nel confronto descrittivo, il set finale mostra risultati complessivamente migliori: GQS più elevato, EBI inferiore e più referti conformi o quasi conformi. Tuttavia, poiché i due dataset sono differenti, non è possibile attribuire con certezza tale miglioramento all'affinamento del modello.

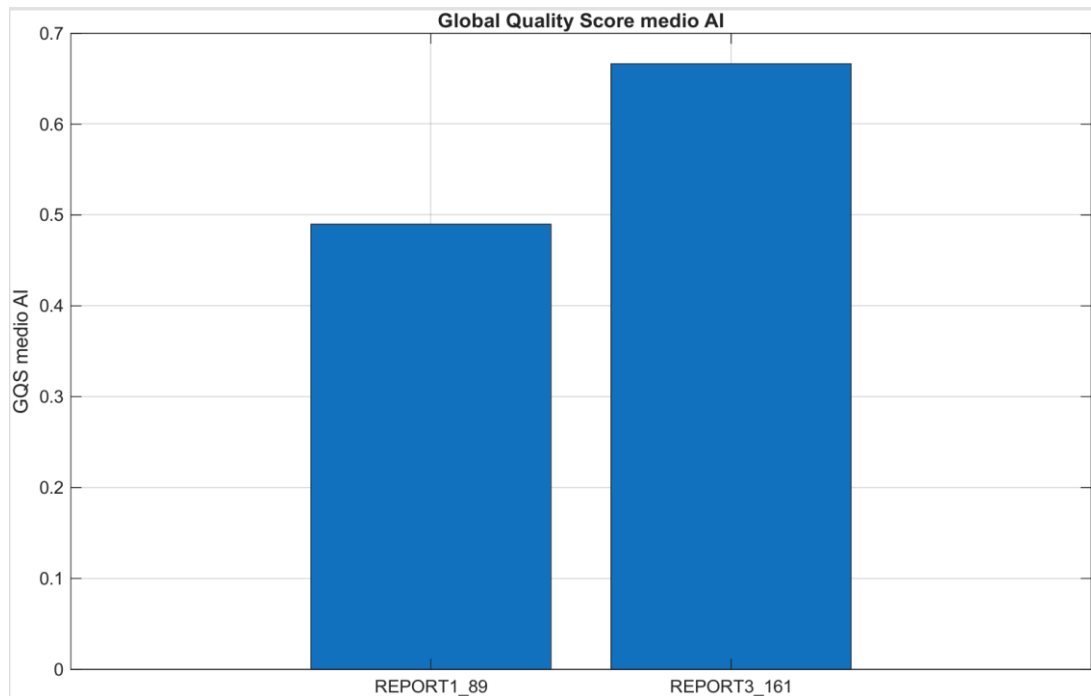


Figura 10: Global Quality Score medio calcolato tramite AI.

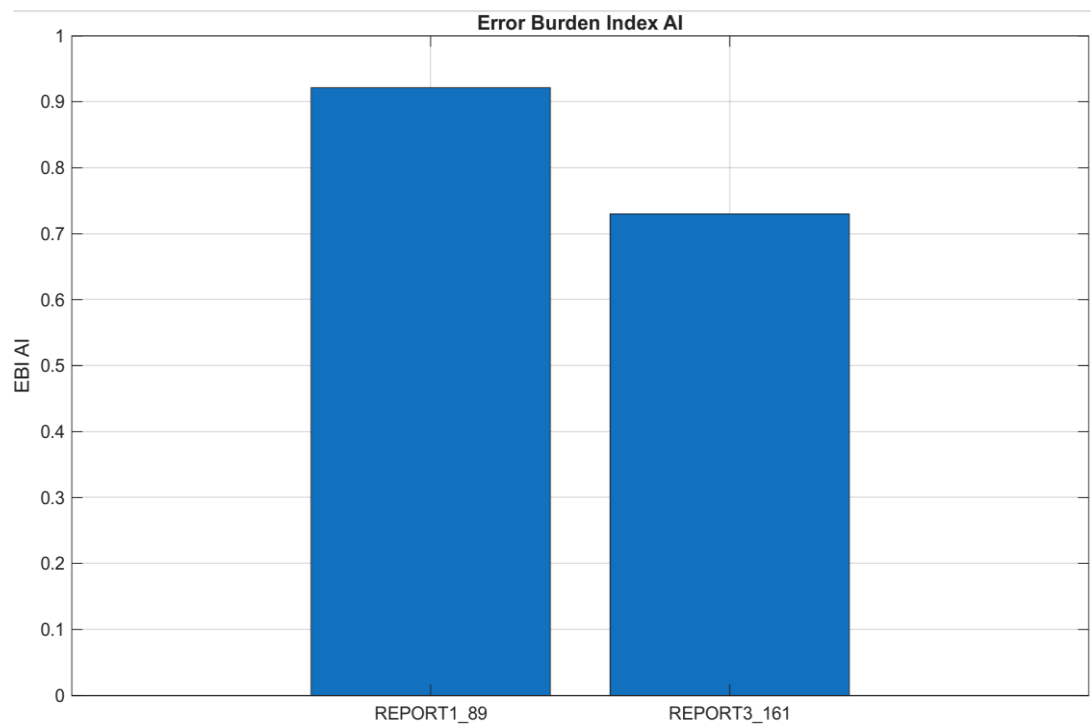


Figura 11: Error Burden Index calcolato tramite AI.

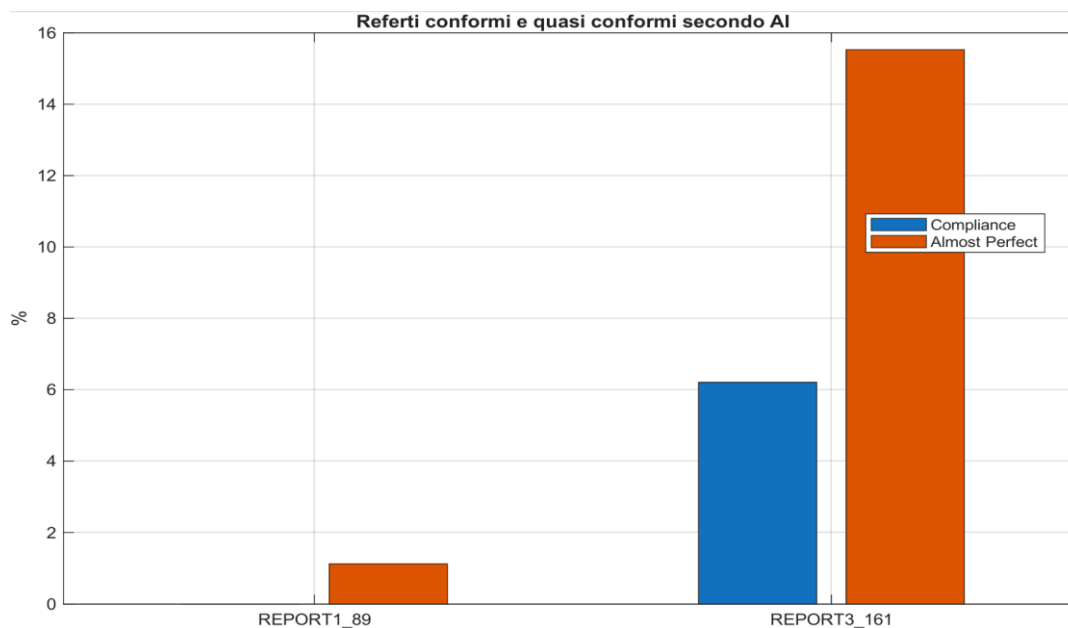


Figura 12: Referti conformi e quasi conformi secondo l'AI.

4.2.3. Concordanza tra valutazione AI e valutazione umana

Sono stati analizzati indicatori di concordanza tra valutazione automatica e valutazione manuale, al fine di verificare in che misura il sistema fosse in grado di riprodurre il giudizio del revisore umano e di individuare eventuali discrepanze nei diversi parametri considerati.

Nel set finale di validazione, la **concordanza tollerante** entro ± 1 livello risulta pari al **97,52%** per la Qualità e al **99,38%** per il Copia-Incolla improprio. Questo indica che anche quando AI e valutatore umano non coincidono esattamente, la discrepanza è nella grande maggioranza dei casi contenuta entro un solo livello della scala.

Per i parametri ordinali, Qualità e Copia-Incolla, è stato calcolato il **Mean Absolute Error** (MAE) con l'obiettivo di quantificare l'entità dello scostamento tra il punteggio assegnato dall'AI e quello attribuito dal valutatore umano. Il MAE esprime la differenza media assoluta tra le due valutazioni: un valore pari a zero indicherebbe una perfetta coincidenza dei punteggi, mentre valori crescenti corrispondono a una maggiore distanza tra giudizio automatico e manuale.

Nel set finale di validazione, il MAE della Qualità è pari a **0,379**, indicando che il punteggio assegnato dall'AI si discosta mediamente da quello umano di meno di mezzo punto su una scala da 1 a 4. Per il Copia-Incolla, il MAE è pari a **0,050**, evidenziando una quasi completa sovrapposizione tra le due valutazioni su una scala da 0 a 2.

Il **Full Match** è risultato pari al **44,72%**. Questo significa che, in quasi la metà dei referti, la valutazione dell'AI coincide integralmente con quella umana su tutti e quattro i criteri considerati: qualità, acronimi/simboli non consentiti, Copia-Incolla improprio e struttura del referto.

La **Critical Discordance** è risultata pari al **27,95%**. Questo indicatore identifica i casi in cui si verifica almeno una discrepanza rilevante, definita come discordanza sulla struttura, discordanza sulla presenza di acronimi o simboli non consentiti oppure differenza di almeno due livelli nel punteggio di qualità. Dal punto di vista pratico, questo indicatore è importante perché consente di distinguere le semplici differenze marginali dalle situazioni in cui AI e valutatore umano esprimono giudizi realmente diversi.

Gli indici di accordo corretti per il caso mostrano risultati differenziati in funzione del tipo di variabile analizzata. Questa analisi è stata introdotta per valutare non soltanto la percentuale di concordanza osservata tra AI e revisore umano, ma anche il livello di accordo al netto della componente che potrebbe verificarsi casualmente. In questo senso, gli indici kappa forniscono una misura più robusta dell'allineamento tra valutazione automatica e valutazione manuale, poiché tengono conto della probabilità di accordo dovuta al caso.

Nel set finale di validazione, il **Cohen's kappa** è risultato pari a **0,654** per la rilevazione di acronimi o simboli non consentiti e pari a **0,776** per la valutazione della struttura del referto. Tali valori indicano un accordo da buono a elevato per le variabili binarie considerate. In particolare, il risultato ottenuto sulla struttura del referto evidenzia un'elevata capacità dell'algoritmo di avvicinarsi al giudizio umano nella distinzione tra referti strutturalmente corretti e non corretti. Anche per la rilevazione di acronimi o simboli non consentiti il valore di kappa suggerisce un buon livello di concordanza, pur lasciando emergere una quota residua di discrepanze tra AI e revisore umano.

Per le variabili ordinali, invece, il **weighted kappa** è risultato pari a **0,283** per la qualità complessiva del referto e pari a **0,418** per il Copia-Incolla improprio. Questi risultati indicano una concordanza corretta per il caso più contenuta rispetto a quanto osservato per le variabili binarie. Tale andamento può essere ricondotto alla maggiore complessità interpretativa delle variabili valutative a più livelli, nelle quali il giudizio non si limita alla semplice distinzione tra presenza e assenza di una criticità, ma richiede una graduazione del livello di qualità o di problematicità rilevato.

Nel complesso, questi risultati evidenziano che l'algoritmo mostra un buon allineamento con il revisore umano quando il criterio di valutazione è binario e maggiormente oggettivabile, come nel caso della struttura del referto e della presenza di acronimi o simboli non consentiti. Al contrario, la concordanza risulta più contenuta per le variabili ordinali, soprattutto per la qualità complessiva, che implica una componente valutativa più articolata. Questo conferma la necessità di interpretare le metriche di accordo in relazione alla natura del parametro analizzato e alla complessità del giudizio richiesto.

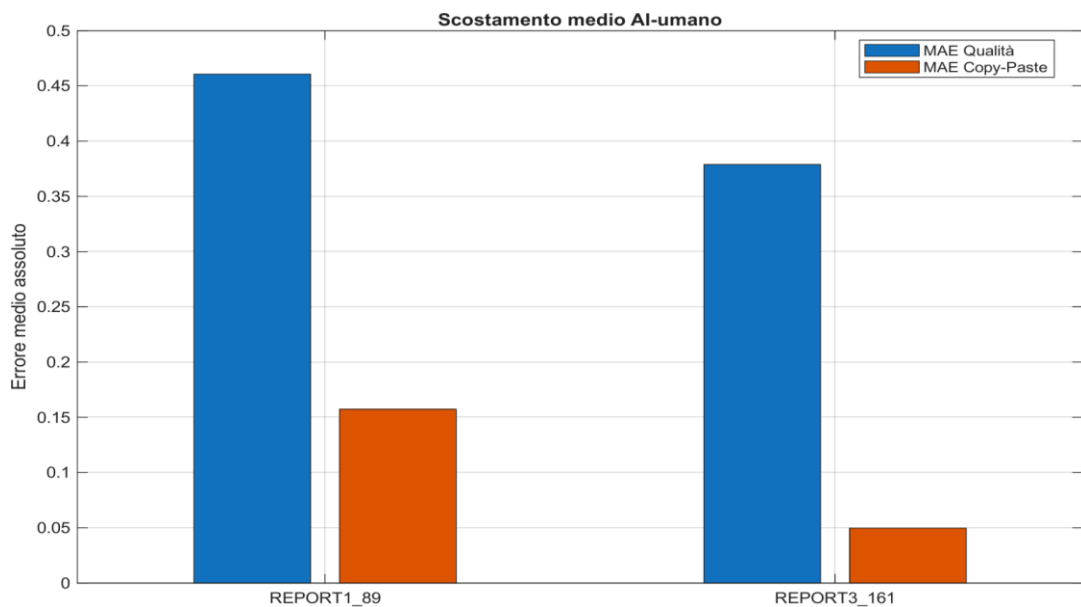


Figura 13: KPI dello scostamento medio tra valutazione AI e valutazione umana.

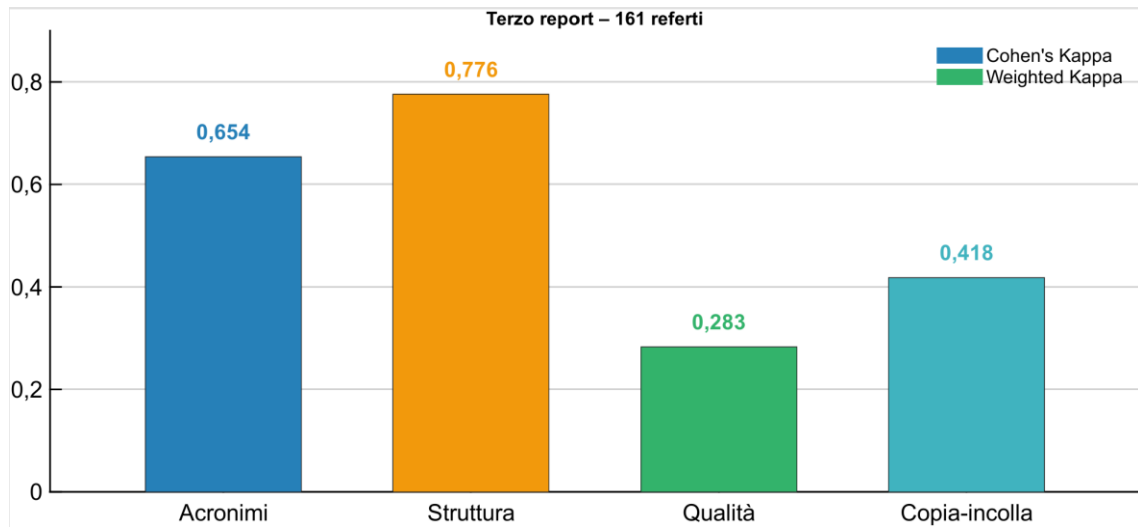


Figura 14: Cohen's Kappa e Weighted Kappa.

4.2.4. KPI organizzativi per l'identificazione dei referti da revisionare

Per valutare il potenziale impatto organizzativo della soluzione AI, i referti sono stati classificati come “da revisionare” nel caso in cui presentassero almeno una non piena conformità rispetto ai criteri considerati: qualità inferiore al punteggio massimo, presenza di acronimi o simboli non consentiti, copy-paste improprio maggiore di zero o struttura non corretta.

Nel set finale di validazione, il valutatore umano classifica come **da revisionare** l'**80,75%** dei referti, mentre l'AI segnala il **93,79%** dei casi. L'algoritmo tende quindi a essere più prudente e a generare un numero maggiore di segnalazioni rispetto al revisore umano.

In termini assoluti, sui 161 referti del set finale:

- 126 sono **veri positivi**, cioè classificati come da revisionare sia dall'AI sia dall'umano;

- 6 sono **veri negativi**, cioè classificati come da non revisionare sia dall'AI sia dall'umano;
- 25 sono **falsi positivi**, cioè classificati dall'AI come da revisionare ma considerati non da revisionare dal valutatore umano;
- 4 sono **falsi negativi**, cioè classificati dall'AI come non da revisionare ma ritenuti da revisionare dal valutatore umano.

La **Triage Accuracy** è pari all'**81,99%**, mentre la **Precision** raggiunge l'**83,44%**. Questo significa che la maggior parte dei referti segnalati dall'AI come meritevoli di revisione risulta effettivamente tale anche secondo la valutazione umana.

Il dato più rilevante dal punto di vista organizzativo è il **Recall**, pari al **96,92%**. L'algoritmo intercetta quindi quasi tutti i referti che il valutatore umano considera da revisionare, lasciandone fuori solo una quota molto ridotta. I **falsi negativi** rappresentano infatti appena il **2,48%** del campione complessivo.

L'**F1-score**, pari all'**89,68%**, conferma un buon equilibrio tra capacità di intercettare i referti da revisionare e affidabilità delle segnalazioni prodotte. La specificità, pari al **19,35%**, rimane invece contenuta: l'AI fatica a riconoscere i referti che non richiedono revisione e tende a segnalarne una parte come potenzialmente problematici.

Nel complesso, i KPI organizzativi suggeriscono che la soluzione AI possiede un elevato potenziale come strumento di pre-screening, in grado di supportare l'identificazione dei referti da controllare e ridurre il rischio di mancata intercettazione di documenti problematici. Tuttavia, la bassa specificità e l'elevato carico di revisione suggerito evidenziano che, allo stato attuale, il sistema non è configurato per sostituire il giudizio umano o per ridurre drasticamente il numero di referti da revisionare, ma per orientare in modo prudentiale l'attenzione del revisore.

Nel Report 1, l'AI classificava come da **revisionare** il **100%** dei referti analizzati. Tale comportamento determinava un **recall** pari al **100%** e l'assenza di falsi negativi, indicando che nessun referto ritenuto problematico dal valutatore umano veniva escluso dalla revisione automatica. Tuttavia, la scelta di segnalare indistintamente tutti i documenti comportava una specificità nulla, poiché il sistema non era in grado di

riconoscere i referti che, secondo il giudizio umano, non necessitavano di approfondimento.

Questo risultato evidenzia come, nella fase preliminare, l'algoritmo adottasse una logica fortemente prudentiale, privilegiando la massima sensibilità a scapito della capacità discriminante. Sebbene tale impostazione riduca il rischio di trascurare referti critici, essa limita l'utilità organizzativa del sistema, poiché non consente di ridurre il carico di lavoro del revisore né di definire una reale priorità tra i documenti.

Il confronto con il Report 3 mostra un'evoluzione del comportamento del modello: pur mantenendo un recall molto elevato, pari al 96,92%, l'AI non segnala più automaticamente tutti i referti come da revisionare. L'introduzione di alcuni veri negativi indica quindi un iniziale miglioramento nella capacità di distinguere i documenti potenzialmente critici da quelli non meritevoli di revisione, anche se la specificità rimane ancora contenuta e suscettibile di ulteriore ottimizzazione.

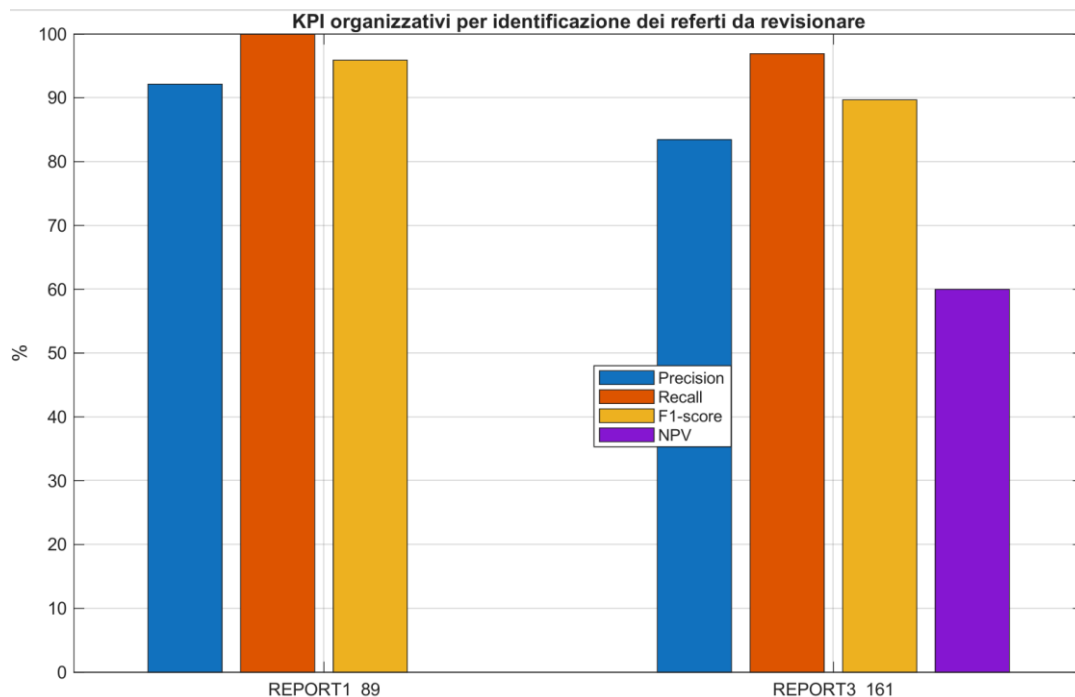


Figura 15: KPI per l'identificazione dei referti da revisionare.

4.6.5. Analisi della direzione dell'errore

È stata infine analizzata la direzione degli scostamenti tra output AI e valutazione umana, con riferimento alla Qualità, al Copia-Incolla improprio e al Global Quality Score.

Nel set finale di validazione, l'AI assegna un punteggio di qualità:

- superiore a quello umano nell'**8,70%** dei casi;
- inferiore a quello umano nel **26,71%** dei casi;
- identico a quello umano nel **64,60%** dei casi.

Il **delta medio della qualità** è pari a **-0,180**, evidenziando una lieve tendenza dell'algoritmo a essere più severo rispetto al valutatore umano nella stima della qualità complessiva del referto.

Per il copia-incolla improprio, l'AI attribuisce un valore

- superiore a quello umano nel **4,35%** dei casi;
- non si osservano casi di sottostima;
- identico a quello umano nel **95,65%** dei casi.

Il **delta medio del copia-incolla** è pari a **0,050**, indicando una lieve tendenza dell'algoritmo a rilevare una criticità leggermente maggiore rispetto al valutatore umano, seppur con intensità molto contenuta. Anche il **delta medio del GQS**, pari a **-0,055**, conferma una tendenza complessiva dell'AI ad attribuire una qualità globale lievemente più bassa rispetto al giudizio umano.

Nel Report 1, la tendenza alla sottostima della qualità risultava più marcata: l'AI assegnava una qualità inferiore rispetto all'umano nel **44,94%** dei casi, con un delta medio pari a **-0,438**. Il delta medio del GQS era pari a **-0,106**.

Nel set finale di validazione, la minore entità del bias negativo sulla qualità e sul GQS suggerisce un comportamento valutativo più vicino a quello umano, pur permanendo una lieve tendenza prudenziale dell'algoritmo.

Questo comportamento è coerente con la tendenza prudenziale già osservata nei KPI organizzativi. L'algoritmo appare infatti orientato a segnalare un numero elevato di referti

potenzialmente critici, riducendo il rischio di falsi negativi ma aumentando al tempo stesso il numero di falsi positivi.

4.7. Sintesi dei risultati principali

I risultati ottenuti evidenziano che il sistema ReportAId mostra un buon potenziale come strumento di supporto al controllo qualità dei referti clinici. Le prestazioni risultano particolarmente solide per l'identificazione dei referti da revisionare, grazie a un recall elevato e a un numero contenuto di falsi negativi. Questo aspetto è rilevante dal punto di vista organizzativo, poiché suggerisce che il sistema possa essere utilizzato come strumento di pre-screening per orientare l'attenzione del revisore umano verso i documenti potenzialmente più critici.

D'altra parte, la bassa specificità e l'elevata percentuale di referti segnalati come da revisionare indicano che il sistema è ancora poco selettivo. In questa configurazione, l'AI non consente quindi una riduzione sostanziale del carico di revisione, ma può contribuire a rendere il processo più strutturato, tracciabile e prudentiale.

Dal punto di vista dei singoli parametri, il Copia-Incolla risulta il criterio più stabile, mentre la Qualità complessiva rimane l'ambito più complesso da automatizzare, a causa della sua natura soggettiva e multidimensionale. Acronimi e Struttura mostrano buoni margini di miglioramento, soprattutto attraverso una definizione più condivisa delle regole e una maggiore integrazione delle specificità redazionali delle diverse specialità cliniche.

Nel complesso, l'analisi conferma che l'intelligenza artificiale può rappresentare un valido supporto al processo di revisione dei referti, ma non può essere considerata, allo stato attuale, un sostituto del giudizio umano. Il suo ruolo più appropriato appare quello di strumento di supporto decisionale, in grado di aumentare la tempestività e la sistematicità del controllo qualità, mantenendo la supervisione finale da parte del personale esperto.

Capitolo 5

Discussione

L'analisi condotta ha permesso di valutare l'impiego di un sistema di intelligenza artificiale basato su Large Language Models come supporto al controllo qualità della refertazione clinica. Il lavoro si è sviluppato su due livelli tra loro complementari. Da un lato, sono stati analizzati i risultati prodotti dal gruppo ReportAId nelle diverse fasi del progetto, osservando l'evoluzione delle prestazioni del sistema rispetto al giudizio umano. Dall'altro, i dati disponibili sono stati rielaborati mediante MATLAB, con l'obiettivo di costruire un insieme strutturato di KPI in grado di descrivere sia la qualità degli output generati dall'AI sia il loro grado di accordo con la valutazione manuale.

L'esame dei report prodotti da ReportAId mostra un progressivo consolidamento del sistema nel corso del progetto. Le prime analisi hanno consentito di individuare le principali aree di disallineamento tra algoritmo e revisione umana, in particolare per i parametri Acronimi e Qualità. Le fasi successive hanno beneficiato di un progressivo affinamento dei criteri di valutazione, della revisione delle regole applicative e dell'aggiornamento delle liste di riferimento, con un miglioramento complessivo delle concordanze osservate. Questo andamento evidenzia come il confronto iterativo con gli esperti clinici rappresenti un passaggio centrale nello sviluppo di strumenti di AI applicati a processi complessi e caratterizzati da una componente interpretativa.

I risultati dei singoli parametri mettono in luce comportamenti differenti del sistema. Il parametro Copia-Incolla risulta generalmente il più stabile, con valori di concordanza

elevati nelle diverse fasi del progetto. Anche il parametro Struttura mostra una buona capacità dell'algoritmo di riconoscere la conformità formale del referto, pur mantenendo alcune criticità nei casi in cui la prassi redazionale clinica si discosta da schemi rigidamente standardizzati. Il parametro Acronimi, inizialmente più problematico, evidenzia un miglioramento significativo nel tempo, confermando l'importanza di disporre di liste condivise e aggiornate di sigle consentite. La Qualità si conferma invece il parametro più complesso, poiché richiede una valutazione complessiva del contenuto del referto e coinvolge aspetti meno immediatamente oggettivabili, come chiarezza, completezza, appropriatezza e coerenza redazionale.

Questa diversa natura dei parametri emerge anche dall'analisi KPI sviluppata sui dataset effettivamente ricostruibili e confrontabili tra AI e valutazione umana. Nel complesso, il sistema mostra risultati più solidi sui criteri maggiormente formalizzabili e più facilmente traducibili in regole di controllo, mentre presenta una maggiore variabilità sui giudizi che richiedono una sintesi interpretativa più articolata. Tale evidenza non rappresenta un limite inatteso, ma riflette la diversa complessità delle dimensioni analizzate.

Particolarmente rilevante è il confronto tra concordanza esatta e concordanza tollerante. Per le variabili ordinali, come Qualità e Copia-Incolla improprio, uno scostamento di un solo livello tra AI e umano può esprimere una differenza contenuta nella valutazione, più che una vera e propria divergenza sostanziale. L'impiego della concordanza tollerante consente quindi di cogliere meglio il grado di prossimità tra i giudizi, evitando di interpretare in modo eccessivamente rigido differenze marginali. Nel caso della Qualità, questo aspetto è particolarmente significativo, poiché la concordanza entro ± 1 livello risulta molto più elevata rispetto alla coincidenza esatta, indicando che il sistema tende spesso a collocarsi vicino al giudizio umano anche quando non assegna lo stesso identico punteggio.

Accanto alle metriche sui singoli parametri, l'introduzione di indicatori complessivi ha permesso di leggere le prestazioni dell'AI in modo più integrato. Il Full Match, ad esempio, misura la coincidenza simultanea tra sistema e valutatore umano su tutte le quattro dimensioni considerate. Si tratta di un indicatore particolarmente stringente, poiché basta una singola discrepanza su uno dei parametri per escludere il referto dalla piena corrispondenza. Il suo valore deve quindi essere interpretato come una misura

prudenziale dell'allineamento complessivo e non come una semplice sintesi delle concordanze parametro per parametro.

Anche il Critical Discordance fornisce un'informazione importante, poiché individua i casi in cui la differenza tra AI e valutazione umana non è marginale, ma riguarda una discrepanza più significativa. Dal punto di vista pratico, questo indicatore può aiutare a comprendere in quali situazioni il sistema richieda maggiore attenzione, cautela interpretativa o ulteriori miglioramenti. La sua introduzione risulta quindi utile non solo per descrivere le prestazioni del modello, ma anche per individuare le aree di maggiore criticità in prospettiva applicativa.

Nel loro insieme, i risultati suggeriscono che ReportAId possa rappresentare uno strumento promettente di supporto al controllo qualità dei referti clinici. Il valore del sistema non deve essere attualmente interpretato in termini di sostituzione del giudizio umano, ma piuttosto nella capacità di affiancare il processo di revisione, fornendo una prima analisi sistematica e riproducibile dei documenti. In un contesto caratterizzato da grandi volumi di referti e da attività di controllo potenzialmente onerose, un sistema di questo tipo potrebbe contribuire a individuare più rapidamente i documenti meritevoli di approfondimento, a rendere più omogenei i criteri di prima valutazione e a favorire una maggiore tracciabilità del processo.

L'interesse organizzativo della soluzione risiede quindi nella possibilità di rendere il controllo qualità più strutturato, esteso e misurabile. L'AI potrebbe costituire un supporto utile per ampliare il numero di referti sottoposti a verifica, ridurre il carico associato a una revisione esclusivamente manuale e generare feedback più tempestivi sulle principali aree di criticità. Tuttavia, tali aspetti devono essere considerati come potenziali ricadute applicative e non come effetti direttamente dimostrati dal presente studio. La misurazione concreta di benefici organizzativi, quali la riduzione dei tempi di revisione, la variazione del carico operativo o l'incremento dei documenti controllati, richiederebbe infatti un'analisi dedicata condotta in condizioni di utilizzo reale.

Lo studio presenta inoltre alcuni limiti metodologici. L'analisi quantitativa sviluppata in MATLAB è stata condotta sui dataset per i quali è stato possibile ricostruire una corrispondenza certa tra output AI e valutazione umana, limitando quindi l'elaborazione approfondita a due set di referti. Inoltre, i dati analizzati appartengono a fasi differenti del

progetto e non rappresentano un unico campione mantenuto costante lungo tutte le iterazioni. Un ulteriore elemento da considerare riguarda la valutazione umana, assunta come riferimento di confronto ma inevitabilmente esposta a una quota di soggettività, soprattutto per parametri interpretativi come la Qualità. Infine, il lavoro si concentra sulla descrizione delle prestazioni del sistema e sul loro possibile significato operativo, senza estendersi alla valutazione di indicatori economici, temporali o di impatto clinico-organizzativo diretto.

Nonostante tali limiti, l'analisi svolta consente di trarre indicazioni utili. Il sistema mostra una buona capacità di affrontare valutazioni formalizzabili e un comportamento generalmente coerente anche su aspetti più complessi. Le discrepanze residue non devono essere considerate soltanto come errori del modello, ma anche come informazioni utili per migliorare le regole di valutazione, perfezionare le liste di riferimento e chiarire ulteriormente i criteri adottati nel controllo qualità della refertazione. In questa prospettiva, l'integrazione dell'AI può essere vista non solo come un supporto tecnico, ma anche come un'occasione per rendere più espliciti, condivisi e verificabili i criteri di qualità documentale.

Capitolo 6

Conclusione

Come detto precedentemente, a partire dai risultati generati dal progetto ReportAId, è stata sviluppata un'analisi quantitativa mediante MATLAB, finalizzata alla costruzione di KPI capaci di descrivere la qualità degli output dell'algoritmo, il grado di concordanza con la valutazione umana e il possibile valore applicativo della soluzione nel processo di controllo qualità.

L'analisi dei report ha evidenziato un progressivo affinamento del sistema lungo le diverse fasi progettuali, favorito dal confronto con gli esperti clinici e dalla revisione dei criteri di valutazione. I parametri più oggettivabili, come Copia-Incolla, Struttura e Acronimi, hanno mostrato livelli di concordanza generalmente elevati o in miglioramento, mentre il parametro Qualità si è confermato il più complesso, in ragione della sua natura più interpretativa e multidimensionale.

La rielaborazione dei dati mediante KPI ha permesso di approfondire ulteriormente il comportamento del sistema, distinguendo tra accordo puntuale, accordo tollerante e corrispondenza complessiva del profilo valutativo. In particolare, gli indicatori aggregati, come Full Match e Critical Discordance, hanno fornito una lettura più completa delle prestazioni dell'AI, evidenziando non solo il livello di allineamento con il giudizio umano, ma anche le aree in cui permane una maggiore distanza tra le due valutazioni.

Nel complesso, i risultati ottenuti supportano l'idea che l'intelligenza artificiale possa rappresentare un valido strumento di affiancamento alla revisione umana dei referti

clinici. Il suo contributo principale non risiede nella sostituzione dell'operatore, ma nella possibilità di realizzare una prima analisi sistematica, riproducibile e scalabile, utile a individuare documenti potenzialmente critici e a rendere il processo di controllo qualità più strutturato.

Il lavoro svolto conferma, inoltre, l'importanza di valutare l'introduzione di tecnologie innovative non solo sulla base delle loro prestazioni tecniche, ma anche attraverso indicatori capaci di descriverne l'utilità all'interno dei processi organizzativi. In questo senso, il framework di KPI sviluppato costituisce uno strumento metodologico utile per interpretare in modo più completo il ruolo dell'AI nella refertazione clinica e per supportare future valutazioni della sua integrazione nei flussi operativi.

Sviluppi successivi potrebbero estendere l'analisi a campioni più ampi e uniformi, prevedere l'impiego di un dataset di riferimento stabile lungo le diverse fasi di affinamento del sistema e includere indicatori di processo, quali tempi di revisione, carico operativo e numero di referti effettivamente controllati. Tali approfondimenti permetterebbero di misurare in modo più diretto l'impatto organizzativo della soluzione e di valutarne l'efficacia in condizioni di utilizzo reale.

In conclusione, l'esperienza analizzata mostra come l'intelligenza artificiale, se inserita in modo controllato e valutata attraverso metriche adeguate, possa offrire un contributo concreto al miglioramento dei processi di controllo della qualità documentale in ambito sanitario. La trasformazione digitale della refertazione non si esaurisce nella produzione di documenti in formato elettronico, ma richiede strumenti capaci di renderne la qualità più osservabile, confrontabile e migliorabile nel tempo

Riferimenti Bibliografici:

- [1] Bilgiç, D., & Camgöz Akdağ, H. (2023). Digital Transformation Readiness Factors in Healthcare. *Hospital topics*, 101(3), 199–207. <https://doi.org/10.1080/00185868.2021.2002745>
- [2] Shah, S. M., & Khan, R. A. (2020). *Secondary Use of Electronic Health Record: Opportunities and Challenges*. arXiv.
- [3] Digital Transformation in Healthcare: 5 Benefits (2023).
- [4] World Health Organization (WHO). (2016). *Electronic Health Records: Manual for Developing Countries*.
- [5] Eur J Public Health. 2024 Oct 28;34(Suppl 3):ckae144.074. Doi: 10.1093/eurpub/ckae144.074)
- [6] Carol Friedman; Thomas C. Rindflesch, “Natural language processing in clinical text,” *Journal of Biomedical Informatics*, 2013.
- [7] Istituto della Enciclopedia Italiana Treccani. *Referto*. Vocabolario Treccani.
- [8] Regione Lombardia – Direzione Generale Salute. *Manuale della Documentazione Sanitaria e Sociosanitaria*. Approvato con D.G.R. n. IX/4659 del 9 gennaio 2013; 2013
- [9] Conferenza Permanente per i Rapporti tra lo Stato, le Regioni e le Province Autonome di Trento e Bolzano. *Linee guida per la dematerializzazione della documentazione clinica in diagnostica per immagini – Normativa e prassi*. Rep. Atti n. 81/CSR; 4 aprile 2012.
- [10] Cappelletti, P. Il “referto” in Medicina di Laboratorio. *Rivista di Medicina di Laboratorio – Journal of Laboratory Medicine*. 2004;5(3):197–208.
- [11] Lukić, V. Laboratory Information System – Where are we Today? *Journal of Medical Biochemistry*. 2017;36(3):220–224. doi: 10.1515/jomb-2017-0021.
- [12] European Society of Radiology (ESR). ESR paper on structured reporting in radiology. *Insights into Imaging*. 2018;9(1):1–7. doi: 10.1007/s13244-017-0588-8.

- [13] Radiological Society of North America (RSNA). *RadLex Radiology Lexicon*. RSNA Practice Tools, Data Tools and Standards.
- [14] Ministero della Salute; Consiglio Superiore di Sanità – Sezione I. *Linee guida: Tracciabilità, raccolta, trasporto, conservazione e archiviazione di cellule e tessuti per indagini diagnostiche di Anatomia Patologica*. Roma: Ministero della Salute; 2015.
- [15] HL7 Italia. *Referto di Specialistica Ambulatoriale – CDA-RSA*. Profilo italiano per la standardizzazione della refertazione di prestazioni ambulatoriali specialistiche.
- [16] Agenzia per l'Italia Digitale (AgID). *Fascicolo Sanitario Elettronico 2.0: Linee guida per l'attuazione*. Roma: AgID; 2022.
- [17] Slawomirski, L., Lindner, L., de Bienassis, K., Haywood, P., Hashiguchi, T. C. O., Steentjes, M., & Oderkirk, J. *Progress on implementing and using electronic health record systems: Developments in OECD countries as of 2021*. OECD Health Working Papers, No. 160. Paris: OECD Publishing; 2023. doi: 10.1787/4f4ce846-en.
- [18] Kim, M. K., Roupheal, C., McMichael, J., Welch, N., & Dasarathy, S. Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation. *Gut and Liver*. 2024;18(2):201–208. doi: 10.5009/gnl230272.
- [19] Kruse, C. S., Smith, B., Vanderlinden, H., & Nealand, A. Security Techniques for the Electronic Health Records. *Journal of Medical Systems*. 2017;41(8):127. doi: 10.1007/s10916-017-0778-4.
- [20] Decreto Legislativo 7 marzo 2005, n. 82. *Codice dell'Amministrazione Digitale (CAD)* e successive modifiche e integrazioni. Gazzetta Ufficiale della Repubblica Italiana, n. 112 del 16 maggio 2005, Suppl. Ord. n. 93.
- [21] Agenzia per l'Italia Digitale (AgID). *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. Roma: AgID; 2021.
- [22] Parlamento Europeo e Consiglio dell'Unione Europea. *Regolamento (UE) 2016/679 del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali e alla libera circolazione di tali dati – General Data Protection Regulation (GDPR)*. Gazzetta Ufficiale dell'Unione Europea; 2016.

- [23] Garante per la Protezione dei Dati Personali. *Chiarimenti sull'applicazione della disciplina per il trattamento dei dati relativi alla salute in ambito sanitario*. Provvedimento n. 55 del 7 marzo 2019.
- [24] Parlamento Europeo e Consiglio dell'Unione Europea. *Regolamento (UE) 2017/745 del 5 aprile 2017 relativo ai dispositivi medici – Medical Device Regulation (MDR)*. Gazzetta Ufficiale dell'Unione Europea; 2017.
- [25] Medical Device Coordination Group (MDCG). *MDCG 2019-11 Rev.1 – Qualification and classification of software: Regulation (EU) 2017/745 and Regulation (EU) 2017/746*. European Commission; June 2025.
- [26] Parlamento Europeo e Consiglio dell'Unione Europea. *Regolamento (UE) 2024/1689 del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale – Artificial Intelligence Act (AI Act)*. Gazzetta Ufficiale dell'Unione Europea; 2024.
- [27] European Commission. *Artificial Intelligence in Healthcare*. Directorate-General for Health and Food Safety.
- [28] Federazione Nazionale degli Ordini dei Medici Chirurghi e degli Odontoiatri (FNOMCeO). *Codice di Deontologia Medica*. Testo approvato nel 2014 con aggiornamenti successivi.
- [29] Dolin, R. H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F. M., Biron, P. V., & Shabo Shvo, A. HL7 Clinical Document Architecture, Release 2. *Journal of the American Medical Informatics Association*. 2006;13(1):30–39. doi: 10.1197/jamia.M1888.
- [30] Health Level Seven International (HL7). *Fast Healthcare Interoperability Resources (FHIR), Release 5.0.0*. Ann Arbor, MI: HL7 International; 2023.
- [31] National Electrical Manufacturers Association (NEMA). *Digital Imaging and Communications in Medicine (DICOM) Standard*. Medical Imaging & Technology Alliance – NEMA. Edizione corrente.
- [32] Torab-Miandoab, A., Samad-Soltani, T., Jodati, A., & Rezaei-Hachesu, P. Interoperability of heterogeneous health information systems: A systematic literature review. *BMC Medical Informatics and Decision Making*. 2023;23:18. doi: 10.1186/s12911-023-02115-5.

- [33] Jendle, J., Adolfsson, P., Choudhary, P., Dove, K., Fleming, A., Klonoff, D. C., Mader, J. K., Oliver, N., Sherr, J. L., Šoupal, J., & Heinemann, L. A narrative commentary about interoperability in medical devices and data used in diabetes therapy from an academic EU/UK/US perspective. *Diabetologia*. 2024;67(2):236–245. doi: 10.1007/s00125-023-06049-5.
- [34] Joint Commission International. *Standards*. Joint Commission International.
- [35] Centro Diagnostico Italiano. *Joint Commission International: settimo rinnovo per il CDI*. 17 febbraio 2025.