Dati in medicina: una guida operativa alla preparazione di dataset per l'intelligenza artificiale

Introduzione: il fondamento dell'IA medica - Il primato dei dati di alta qualità

Nel campo dell'Intelligenza Artificiale applicata alla medicina, vige un principio tanto semplice quanto inesorabile: *Garbage In, Garbage Out*, ovvero "spazzatura in ingresso, spazzatura in uscita". Un modello algoritmico, per quanto sofisticato, se addestrato su dati imprecisi, incompleti o distorti, produrrà risultati non solo inutili, ma potenzialmente pericolosi per la pratica clinica. La preparazione di un dataset non è, quindi, un mero compito tecnico da relegare agli specialisti informatici; è un pilastro fondamentale della sicurezza del paziente, dell'integrità della ricerca e della validità di ogni futura applicazione clinica dell'IA.

Il percorso di un'informazione medica, dal suo stato fisico a quello di dato digitale, è irto di potenziali insidie. Si pensi a una semplice misurazione della pressione arteriosa: nasce come segnale fisico (la pressione nelle arterie), viene trasdotta da un dispositivo in un valore numerico, trascritta (manualmente o automaticamente) in una cartella clinica elettronica, e infine estratta per essere inclusa in un dataset di ricerca. In ogni passaggio di questa catena, possono essere introdotti errori, bias e incongruenze, trasformando un segnale accurato in un dato fuorviante.

Per questo motivo, la creazione di un dataset medico di alta qualità non è un'impresa solitaria per un *Data Scientist*. Al contrario, richiede fin dall'inizio un mandato interdisciplinare. È un lavoro di squadra che deve coinvolgere: i **clinici**, per definire il quesito di ricerca e validare la plausibilità dei dati; i **biostatistici**, per disegnare lo studio e pianificare l'analisi; gli **esperti legali**, per garantire la conformità normativa, in particolare con il GDPR; e gli **specialisti IT**, per la gestione sicura e l'estrazione dei dati. Solo attraverso questa sinergia di competenze è possibile costruire fondamenta solide su cui edificare algoritmi di IA affidabili e di reale valore clinico.

Parte 1: Il progetto strategico - Pianificare e disegnare un dataset medico

La qualità di un dataset viene determinata molto prima che il primo dato venga raccolto. Essa è il risultato diretto di una pianificazione rigorosa, proattiva e meticolosa. Un errore in questa fase iniziale può generare difetti strutturali che nessuna tecnica di "pulizia" successiva potrà mai correggere del tutto.

1.1 Definire il quesito clinico: dall'ipotesi agli endpoint

Il documento cardine di ogni progetto di ricerca è il **piano di indagine clinica**, o protocollo. Esso rappresenta la "costituzione" dello studio, il testo fondamentale che governa ogni fase successiva, dalla raccolta dati all'analisi finale. All'interno del protocollo, il primo passo è la formulazione di un quesito di ricerca chiaro e preciso. Una domanda generica come "Possiamo predire la sepsi?" deve essere affinata in un obiettivo specifico, misurabile, raggiungibile, rilevante e definito nel tempo (*SMART*).

Una volta definito l'obiettivo, è importante stabilire gli **endpoint**, ovvero gli esiti (outcome) che il modello di IA dovrà predire. Questi devono essere risultati specifici e misurabili, come la mortalità a 30 giorni, la durata della degenza in terapia intensiva, o il valore di un particolare esame di laboratorio a una data ora. La scelta degli endpoint primari e secondari non è un dettaglio tecnico, ma una decisione strategica che condiziona l'intera architettura della raccolta dati.

1.2 Disegno dello studio: la scelta fondamentale

Una delle decisioni più impattanti sulla qualità finale del dataset riguarda il disegno dello studio. Le due opzioni principali, studi prospettici e retrospettivi, presentano compromessi radicalmente diversi.

- **Studio Prospettico**: I dati vengono raccolti *ad hoc* per lo studio in corso. Questo approccio permette un elevato controllo sulla qualità: si possono standardizzare le metodologie di raccolta, definire a priori quali variabili raccogliere e minimizzare la quantità di dati mancanti. Tuttavia, è un processo lungo e costoso.
- **Studio Retrospettivo**: I dati vengono estratti da fonti preesistenti, come le cartelle cliniche elettroniche. Questo metodo è più rapido ed economico, ma introduce problemi

intrinseci. I dati, infatti, sono stati raccolti per finalità di cura clinica, non di ricerca, e sono quindi spesso disomogenei, privi di standardizzazione e afflitti da una grande quantità di valori mancanti e incongruenze.

La scelta tra questi due approcci non è neutrale. Uno studio retrospettivo, pur essendo più accessibile, genera un dataset con difetti congeniti che richiederanno un enorme sforzo di pulizia e, anche dopo tale sforzo, potrebbero persistere bias nascosti. Uno studio prospettico, invece, "progetta la qualità" fin dall'inizio, prevenendo a monte molti dei problemi che affliggono i dati raccolti retrospettivamente, ma ha bisogno di molto tempo per dispiegarsi. La tecnica di preparazione del dato più efficace, in ultima analisi, è un protocollo di studio ben disegnato.

Tabella 1: Confronto tra metodi di raccolta dati (prospettico vs. retrospettivo)

Caratteristica	Disegno Prospettico	Disegno Retrospettivo
Controllo Qualità Dati	Elevato, definito a priori dal protocollo	Basso, dipendente dalla qualità della documentazione clinica preesistente
Standardizzazione	Alta, le procedure di raccolta sono uniformi	Bassa, i dati sono raccolti da operatori diversi in contesti diversi
Dati Mancanti	Ridotti al minimo, la raccolta è mirata	Elevati, molte variabili di interesse per la ricerca non sono di routine clinica
Costo	Elevato	Basso
Tempo	Lungo	Breve
Potenziale di Bias	Basso (se ben disegnato)	Alto (bias di selezione, bias storico)

Idoneità per IA	Ideale, dati puliti e mirati	Richiede un intenso lavoro di pre-processing e pulizia

1.3 Il Piano di gestione dei dati (DMP): il regolamento del vostro dato

Se il protocollo è la costituzione, il **Data Management Plan (DMP)** è il codice di attuazione. È un documento formale che descrive in dettaglio l'intero ciclo di vita del dato, dall'acquisizione alla conservazione finale. Un DMP ben strutturato deve includere:

- **Fonti dei dati**: un elenco esplicito di tutte le fonti che verranno utilizzate (es. EHR, sistemi di laboratorio, archivi di immagini, dati riportati dai pazienti).
- **Ruoli e responsabilità**: una chiara definizione di chi è responsabile per l'inserimento, il controllo qualità, la pulizia e la gestione del database.
- **Procedure di raccolta e gestione**: la specificazione degli strumenti, come le schede di raccolta dati elettroniche, che standardizzano l'inserimento e riducono gli errori.
- **Procedure di controllo qualità**: la predeterminazione dei controlli di validità (es. "edit checks") che verranno applicati per identificare dati errati o anomali in tempo reale.

1.4 L'Iter etico e normativo: consenso, GDPR e governance

La gestione dei dati medici è strettamente regolamentata per proteggere i diritti e la privacy dei pazienti.

- **Consenso informato**: è fondamentale distinguere tra il consenso al trattamento medico e il consenso specifico per l'utilizzo dei dati a fini di ricerca. I pazienti devono essere informati in modo chiaro e completo su come i loro dati verranno utilizzati.
- **GDPR in sanità**: il Regolamento Generale sulla Protezione dei Dati (GDPR) definisce i dati relativi alla salute come "categorie particolari di dati personali", che richiedono un livello di protezione rafforzato. Il loro trattamento è lecito solo in presenza di specifiche basi giuridiche. Per la ricerca scientifica, oltre al consenso, la base giuridica può essere "l'esecuzione di un compito di interesse pubblico" o "motivi di interesse pubblico nel settore della sanità pubblica", a condizione che siano adottate garanzie adeguate per i diritti degli interessati.
- **Comitati etici**: ogni studio deve essere approvato da un comitato etico indipendente, che valuta il protocollo per garantire che la ricerca sia condotta in modo etico, che i rischi per i pazienti siano minimizzati e che i loro diritti siano tutelati.

Parte 2: Guida passo-passo alla preparazione dei dati

Una volta completata la fase strategica, inizia il lavoro tecnico di trasformazione dei dati grezzi in un dataset strutturato, pulito e pronto per essere utilizzato dagli algoritmi di IA.

2.1 Raccolta e integrazione dei dati: assemblare il puzzle

I dati clinici sono tipicamente frammentati in sistemi eterogenei e isolati. Un compito importante è integrarli in un unico repository coerente. Le fonti principali includono:

- **Cartelle cliniche elettroniche (EHR)**: Contengono una ricchezza di informazioni, ma spesso in formati non strutturati (es. note cliniche in testo libero) e con terminologie non standardizzate.
- **Sistemi informativi di laboratorio (LIMS)**: Forniscono dati strutturati, ma richiedono una mappatura dei codici dei test (es. verso standard come LOINC) per essere confrontabili.
- **Sistemi di archiviazione e comunicazione di immagini (PACS)**: Contengono immagini diagnostiche in formato DICOM, che include metadati dettagliati sul paziente e sull'acquisizione che devono essere accuratamente de-identificati.
- **Dati riportati dai pazienti (ePROs)**: Dati raccolti direttamente dai pazienti tramite app o questionari, che offrono una prospettiva unica ma possono introdurre soggettività.

L'uso di un **clinical data warehouse** (magazzino di dati clinici) può facilitare questo processo, fornendo una piattaforma centralizzata per armonizzare e organizzare queste fonti disparate.

2.2 Pulizia dei dati

Questa fase, nota anche come "data cleaning" o "data wrangling", è un processo meticoloso per correggere le imperfezioni dei dati grezzi.

- **Rilevamento e correzione di errori**: identificazione e rettifica di errori di battitura, valori impossibili (es. una frequenza cardiaca negativa) o palesemente errati.
- **Standardizzazione**: un passo critico per garantire la coerenza. Include la conversione di unità di misura (es. tutti i pesi in kg), la mappatura di termini clinici a ontologie standard (es. SNOMED CT per le diagnosi) e l'uniformazione dei formati (es. date e ore).
- **Gestione degli outlier**: gli outlier sono valori che si discostano significativamente dal resto dei dati. È fondamentale distinguere tra un errore di inserimento e un valore

estremo ma clinicamente plausibile (es. un valore di laboratorio critico). Questa distinzione non può essere puramente statistica, ma richiede la supervisione di un esperto clinico.

- **Deduplicazione**: identificazione e fusione di record duplicati per lo stesso paziente, un problema comune quando si integrano dati da più sistemi.

2.3 Gestire il vuoto: un approfondimento sui dati mancanti

I dati mancanti non sono un semplice fastidio; sono una potenziale fonte di bias grave che può invalidare i risultati di uno studio. Ignorarli o gestirli in modo inappropriato può portare a conclusioni errate e pericolose. Il primo passo non è scegliere un algoritmo per "riempirli", ma formulare un'ipotesi clinica sul *perché* mancano.

Ad esempio, in uno studio su un farmaco antidolorifico, se i pazienti con il dolore più intenso non riescono a compilare il diario giornaliero del dolore, questi dati sono "mancanti non a caso" (MNAR). Se si eliminano questi record o li si sostituisce con un valore medio, si sottostimerà sistematicamente il fallimento del farmaco nel controllare il dolore severo, portando alla conclusione errata che il farmaco sia più efficace di quanto non sia in realtà. La gestione dei dati mancanti è, quindi, un compito diagnostico che richiede una stretta collaborazione tra data scientist e clinico.

Tabella 2: Tassonomia dei dati mancanti

Tipo	Definizione	Esempio medico	Rischio principale	Strategia di gestione adeguata
MCAR (Missing Completely at Random)	La probabilità che un dato manchi è indipendente sia dai valori osservati che da quelli non	Un macchinario di laboratorio si rompe e perde alcuni campioni a caso.	Basso. Perdita di potenza statistica.	Eliminazione (se <5% dei dati), imputazione semplice o multipla.

	osservati.			
MAR (Missing at Random)	La probabilità che un dato manchi dipende da altre variabili osservate nel dataset, ma non dal valore mancante stesso.	I pazienti maschi sono meno propensi a compilare un questionario sulla depressione (la "mancanza" dipende dal sesso, che è osservato).	Bias se gestito in modo ingenuo (es. eliminazione).	Imputazione basata su modelli (es. regressione), imputazione multipla.
MNAR (Missing Not at Random)	La probabilità che un dato manchi dipende dal valore stesso della variabile mancante.	Pazienti con reddito molto alto o molto basso omettono di dichiararlo. Pazienti molto malati saltano le visite di follow-up.	Bias grave e sistematico, difficile da correggere.	Richiede modelli statistici avanzati e assunzioni esplicite sul meccanismo di mancanza. L'eliminazione o l'imputazione semplice sono quasi sempre errate.

Le strategie per gestire i dati mancanti vanno dalla semplice **eliminazione** delle righe con valori mancanti (accettabile solo per dati MCAR e in piccola percentuale), all'**imputazione singola** (sostituzione con media, mediana, ecc.), fino a metodi più avanzati. L'**imputazione multipla** è considerata il gold standard: crea diverse versioni complete del dataset, esegue l'analisi su ciascuna e poi combina i risultati, tenendo conto dell'incertezza introdotta dall'imputazione stessa.

2.4 Trasformazione dei dati e feature engineering

Gli algoritmi di machine learning richiedono input numerici e ben formattati. Questa fase adatta i dati puliti per l'addestramento del modello.

- **Codifica**: Conversione di variabili categoriche (es. 'Maschio', 'Femmina') in rappresentazioni numeriche (es. 0, 1).
- **Normalizzazione/Scalatura**: Ridimensionamento delle variabili numeriche a un intervallo comune (es. da 0 a 1) per evitare che le variabili con scale più ampie (es. il costo di una prestazione) dominino il modello rispetto a quelle con scale più piccole (es. un valore di laboratorio).
- **Creazione di feature (feature engineering)**: Derivazione di nuove variabili informative da quelle esistenti (es. calcolo dell'Indice di Massa Corporea o BMI da altezza e peso).

2.5 Divisione del dataset: addestramento, validazione e test

Per valutare onestamente le prestazioni di un modello, è essenziale testarlo su dati che non ha mai visto durante la fase di addestramento. Per questo, il dataset viene suddiviso in tre parti distinte:

- **Training Set**: La porzione più grande dei dati (solitamente 60-80%), utilizzata per "insegnare" al modello a riconoscere i pattern.
- **Validation Set**: Una porzione più piccola (10-20%), utilizzata per ottimizzare i parametri del modello e per valutare l'impatto dell'overfitting (quando il modello impara a memoria i dati di training ma non generalizza bene a nuovi dati).
- **Test Set**: L'ultima porzione (10-20%), che viene "messa da parte" e utilizzata una sola volta, alla fine del processo, per ottenere una stima finale e imparziale delle prestazioni del modello nel mondo reale.

Per dataset di dimensioni ridotte, una tecnica più robusta è la **validazione incrociata (cross-validation)**, che suddivide ripetutamente il dataset in training e validation set per ottenere una stima più stabile delle performance.

Parte 3: Garantire fiducia e conformità - validazione e privacy

Questa sezione si concentra sui processi formali per verificare la qualità dei dati e assicurare che la loro gestione sia conforme alle rigide normative sulla privacy.

3.1 Assicurazione della qualità e validazione dei dati

Oltre alla pulizia ad-hoc, è necessario implementare un framework sistematico per il controllo della qualità, basato su dimensioni ben definite:

- **Completezza**: verifica che tutti i dati richiesti siano presenti.
- **Conformità**: assicura che i dati rispettino i formati e gli standard predefiniti (es. tipi di dato, intervalli di valori consentiti).
- **Plausibilità**: valuta se i dati sono credibili nel contesto del mondo reale (es. l'età di un paziente non può diminuire tra una visita e l'altra).

Il **Data Validation Plan (DVP)**, definito nella fase di pianificazione, viene qui messo in pratica. Esso specifica gli "edit checks" automatici che, ad esempio in una eCRF, possono segnalare errori durante l'inserimento dei dati, migliorando la qualità in tempo reale. Un altro elemento da implementare è l'**audit trail**, un registro immutabile di ogni modifica apportata ai dati (chi, cosa, quando e perché). Questo è un requisito fondamentale per l'integrità dei dati.

3.2 Proteggere l'identità del paziente: anonimizzazione e pseudonimizzazione

La distinzione tra anonimizzazione e pseudonimizzazione è un cardine della conformità al GDPR e presenta implicazioni legali e tecniche profonde.

- **Pseudonimizzazione**: sostituisce gli identificatori diretti (es. nome, codice fiscale) con uno pseudonimo o un codice. Esiste una "chiave" separata e conservata in modo sicuro che permette di re-identificare il soggetto. Secondo il GDPR, i dati pseudonimizzati sono ancora considerati **dati personali** e rimangono soggetti a tutte le tutele del regolamento. È una tecnica comune negli studi clinici, dove la re-identificazione può essere necessaria per motivi di sicurezza del paziente.

- **Anonimizzazione**: È un processo di de-identificazione **irreversibile**. Tutte le informazioni che potrebbero permettere di re-identificare un individuo, anche indirettamente, vengono rimosse o modificate in modo tale che non sia ragionevolmente possibile risalire al soggetto. I dati anonimizzati **non sono più considerati dati personali** ai sensi del GDPR e possono essere utilizzati con minori restrizioni.

La scelta tra queste due tecniche implica un compromesso fondamentale tra la protezione della privacy e l'utilità scientifica dei dati. Tecniche di anonimizzazione robuste, come l'aggregazione dei dati in fasce (es. età 40-50 anni invece dell'età esatta) o l'aggiunta di "rumore" statistico, riducono il rischio di re-identificazione ma allo stesso tempo degradano la qualità e la granularità del dato. Questa perdita di precisione può rendere il dataset inadatto per addestrare modelli di IA che necessitano di dati dettagliati per apprendere correlazioni sottili. Pertanto, la scelta della tecnica non è puramente tecnica o legale, ma strategica: deve essere ponderata in base al quesito di ricerca, documentata e giustificata attraverso un'analisi del rischio, in linea con il principio di accountability del GDPR.

Tabella 3: Anonimizzazione vs. Pseudonimizzazione secondo il GDPR

	Pseudonimizzazione	**Anonimizzazione**
Definizione	Trattamento che impedisce l'attribuzione diretta a un interessato senza informazioni aggiuntive (la "chiave").	Processo che rimuove gli identificatori in modo irreversibile, rendendo l'interessato non più identificabile.
Reversibilità	Reversibile. È possibile risalire all'identità tramite la chiave di decodifica.	Irreversibile. La re-identificazione non è ragionevolmente possibile.
Status Legale (GDPR)	Rimane un dato personale, soggetto al GDPR.	Non è più un dato personale, esce dal campo di applicazione del GDPR.
Esempio di Tecnica	Sostituzione del codice fiscale con un codice alfanumerico.	Aggregazione, randomizzazione, k-anonimato.

Caso d'Uso in Ricerca	Studi clinici in cui è necessario poter ricontattare il paziente o collegare dati longitudinali.	Rilascio di dataset pubblici per analisi statistiche aggregate.
Responsabilità Titolare	Deve proteggere sia i dati pseudonimizzati sia la chiave di decodifica con misure di sicurezza adeguate.	Deve dimostrare (accountability) che il processo di anonimizzazione è robusto e il rischio di re-identificazione è nullo o minimo.

Le tecniche di anonimizzazione includono la **randomizzazione** (aggiunta di rumore statistico) e la **generalizzazione** (riduzione della granularità). Un'altra tecnica chiave di generalizzazione è la **k-anonimità**, che garantisce che ogni record nel dataset sia indistinguibile da almeno k-1 altri record, rendendo difficile l'isolamento di un singolo individuo. Tuttavia, è fondamentale essere consapevoli del rischio di "attacco di collegamento" (linkage attack), in cui un dataset apparentemente anonimo, se combinato con altre fonti di dati pubbliche, può portare alla re-identificazione di individui.

Parte 4: Documentare per la trasparenza - Il ruolo delle Data Cards

Una documentazione chiara, strutturata e trasparente è una pratica scientifica essenziale. Permette la riproducibilità, favorisce un uso responsabile dei dati e costruisce fiducia nella comunità scientifica e nel pubblico.

4.1 La logica della documentazione strutturata

Un semplice file "*README.md*" con una breve descrizione non è sufficiente per i dataset medici, dove la posta in gioco è alta. È necessaria una documentazione standardizzata per comunicare in modo efficace il contesto, i punti di forza e le limitazioni di un dataset a un pubblico eterogeneo (creatori, utenti, revisori, enti regolatori). Questo approccio promuove i principi **FAIR** (Findable, Accessible, Interoperable, Reusable) e costituisce il fondamento dell'**IA responsabile**, consentendo agli utenti di comprendere potenziali bias, considerazioni etiche e casi d'uso appropriati.

Una Data Card non è un semplice inventario del contenuto finale di un dataset. È la narrazione dell'intero ciclo di vita dei dati, un documento che cattura le decisioni critiche, i compromessi accettati e le limitazioni intrinseche del processo. Ad esempio, sapere che un dataset proviene da un unico ospedale universitario avverte immediatamente l'utente che un modello addestrato su di esso potrebbe non generalizzare bene a contesti diversi. Questa narrazione trasforma la documentazione da un elenco statico a un vero e proprio "articolo scientifico" sul dataset stesso, promuovendo la riproducibilità.

4.2 Anatomia di una Data Card

Ispirandosi a framework consolidati come "Datasheets for Datasets" e il "Data Cards Playbook" (vedi Rifermineti alla fine del documento), una Data Card ben strutturata dovrebbe includere sezioni che rispondono a domande fondamentali:

- **Motivazione e uso previsto**: perché il dataset è stato creato? Per quali scopi è adatto e per quali non lo è?
- **Composizione e provenienza**: cosa contiene il dataset? Da dove provengono i dati? Quali erano i criteri di inclusione/esclusione?
- **Processo di raccolta**: come sono stati raccolti i dati? È stato ottenuto il consenso?

- **Pre-elaborazione, trasformazione ed etichettatura**: quali passaggi di pulizia sono stati eseguiti? Come sono state generate le etichette (es. da annotatori umani)?
- **Distribuzione e manutenzione**: come si accede ai dati? Esistono diverse versioni? Verrà aggiornato?
- **Considerazioni etiche e sui bias**: discussione delle limitazioni note, della rappresentatività demografica e dei rischi per la privacy.

4.3 Un template per i dataset medici

Adattando i principi generali al contesto medico, si può definire un template specifico che includa campi cruciali per questo dominio.

Tabella 4: Template per Data Card di un dataset medico

	Campi Chiave e Domande Guida
1. Panoramica del dataset	- *Nome del dataset* (Acronimo): - *Riepilogo* (max 200 parole): descrivere il contenuto, la fonte, la motivazione e i casi d'uso adatti. - *Parole chiave*: (es. Radiologia, Cardiologia, COVID-19).
2. Autori e finanziamenti	- *Autori del dataset*: nomi, affiliazioni. - *Contatto di riferimento*: email o sito web per domande. - *Ente finanziatore*: (es. Ministero della Salute, Fondazione di Ricerca). - *Citazione consigliata*: come citare il dataset nelle pubblicazioni.
3. Motivazione e uso previsto	- *Quesito di ricerca*: quale problema clinico o scientifico si intende risolvere? - *Casi d'uso idonei*: (es. Sviluppo di modelli diagnostici, analisi epidemiologiche) - *Casi d'uso non idonei/fuori scopo*: (es. Non utilizzare per decisioni cliniche dirette senza validazione)

4. Approvazione etica e normativa	- *Comitato etico*: nome del comitato che ha approvato lo studio. - *Numero di protocollo/approvazione*: riferimento ufficiale dell'approvazione. - *Consenso informato*: tipo di consenso ottenuto (es. specifico per la ricerca, deroga del comitato etico).
5. Descrizione della coorte di pazienti	- *Criteri di inclusione/esclusione*: elenco dettagliato dei criteri usati per selezionare i pazienti. - *Demografia della coorte*: statistiche aggregate (es. distribuzione di età, sesso, etnia). - *Periodo di raccolta*: data di inizio e fine della raccolta dati.
6. Provenienza e raccolta dati	- *Fonte dei dati*: (es. EHR dell'Ospedale X, Registro di patologia Y). - *Metodo di raccolta*: prospettico o retrospettivo. - De-identificazione: descrivere il metodo usato (pseudonimizzazione o anonimizzazione) e i rischi residui.
7. Pre-elaborazione ed etichettatura	- *Procedure di Pulizia*: sintesi dei passaggi di pulizia eseguiti (gestione dei dati mancanti, standardizzazione, ecc.) - *Protocollo di etichettatura/annotazione*: se applicabile (es. per immagini), descrivere chi ha etichettato i dati (es. radiologi certificati), quali istruzioni ha seguito e la concordanza tra annotatori.
8. Struttura e campi del dataset	- *Formato dei file*: (es. CSV, DICOM, JSON). - *Dizionario dei dati*: descrizione dettagliata di ogni colonna/variabile, incluse unità di misura e valori possibili. - *Esempio di Record*: un record anonimizzato di esempio.
9. Bias e limitazioni	- *Bias di selezione*: la coorte è rappresentativa della popolazione generale? (es. dati da un solo centro

note	specialistico). - *Bias di misurazione*: potenziali imprecisioni sistematiche negli strumenti o nelle procedure. - *Limitazioni conosciute*: qualsiasi altro fattore che potrebbe influenzare l'interpretazione dei risultati.
10. Accesso e manutenzione	- *Modalità di accesso*: (es. Download pubblico, richiesta formale, ambiente di ricerca sicuro). - *Licenza d'uso*: (es. CC-BY, solo per ricerca non commerciale) - *Controllo delle versioni*: Il dataset è versionato? Esiste un changelog?

4.4 Caso di studio: decostruire la Data Card di CheXpert

Per rendere concreto il concetto, si può analizzare un esempio reale come **CheXpert**, un dataset pubblico su larga scala di radiografie del torace. Popolando il nostro template medico con le informazioni disponibili pubblicamente, si otterrebbe:

- **Panoramica**: Dataset di 224.316 radiografie del torace di 65.240 pazienti, raccolte presso lo Stanford Health Care tra il 2002 e il 2017.
- **Motivazione**: Fornire un grande dataset etichettato per lo sviluppo e la validazione di algoritmi di interpretazione automatica delle radiografie del torace.
- **Provenienza e raccolta**: Dati retrospettivi dall'archivio PACS di Stanford.
- **Etichettatura**: Le etichette per 14 osservazioni radiologiche comuni (es. cardiomegalia, versamento pleurico) sono state estratte automaticamente dai referti radiologici testuali associati, utilizzando un estrattore basato su Natural Language Processing.
- **Bias e limitazioni**: I dati provengono da un singolo centro accademico statunitense, il che potrebbe limitare la generalizzabilità dei modelli. Le etichette generate via NLP possono contenere errori rispetto a una revisione manuale da parte di radiologi esperti.

Questo esercizio dimostra come una Data Card possa riassumere in modo trasparente le informazioni per un utilizzo consapevole del dataset.

Conclusioni

5.1 Dal dato alla diagnosi: chiudere il cerchio

Il processo meticoloso e a tratti arduo di preparazione dei dati non è fine a se stesso. Ogni passaggio, dalla definizione del protocollo alla compilazione della Data Card, è un anello di una catena che collega un'osservazione clinica grezza a un potenziale strumento di IA in grado di migliorare la pratica medica. La fiducia che i clinici, i pazienti e gli enti regolatori riporranno in questi strumenti dipende interamente dalla trasparenza e dal rigore con cui sono stati creati i dati su cui si basano.

5.2 Il paesaggio in evoluzione dei dati medici

Il campo della gestione dei dati medici è in continua e rapida evoluzione. Per gli studenti che oggi si affacciano a queste discipline, è fondamentale essere consapevoli delle tendenze emergenti che plasmeranno il futuro:

- **Dati sintetici**: dati generati artificialmente da modelli di IA che mimano le proprietà statistiche dei dati reali. Possono essere utilizzati per aumentare le dimensioni dei dataset o per condividere dati senza rischi per la privacy, sebbene la loro validazione sia ancora un'area di ricerca attiva.
- **Apprendimento federato (federated learning)**: una tecnica che permette di addestrare modelli di IA su dati distribuiti in più ospedali o centri di ricerca senza che i dati grezzi lascino mai la loro sede originale. Questo approccio permetterebbe di rivoluzionare la collaborazione scientifica, superando molti ostacoli legati alla privacy e alla governance dei dati.
- **Dati multimodali**: la crescente capacità di integrare tipi di dati estremamente diversi — genomica, immagini mediche, note cliniche, dati da sensori indossabili — per creare una visione olistica e multidimensionale del paziente. Questo aprirà la strada a una medicina veramente personalizzata e predittiva.

Affrontare queste nuove frontiere richiederà una comprensione ancora più profonda dei principi fondamentali di qualità, gestione e governance dei dati qui esposti. La competenza nella preparazione di dataset di alta qualità non sarà più una specializzazione di nicchia, ma una competenza di base per ogni futuro professionista della sanità nell'era dell'Intelligenza Artificiale.

Bibliografia

1. Sicurezza dei dati: pseudonimizzazione o anonimizzazione? - Accademia Italiana Privacy, accesso eseguito il giorno settembre 19, 2025, https://www.accademiaitalianaprivacy.it/dettaglioNews.asp?id=646
2. Codice di condotta per l'utilizzo di dati sulla salute a fini didattici e di pubblicazione scientifica - Garante Privacy, accesso eseguito il giorno settembre 19, 2025, https://www.garanteprivacy.it/garante/document?ID=9535402
3. Datasheets for Datasets - arXiv, accesso eseguito il giorno settembre 19, 2025, https://arxiv.org/pdf/1803.09010
4. The Data Cards Playbook - Google Research, accesso eseguito il giorno settembre 19, 2025, https://sites.research.google/datacardsplaybook/
5. Dataset Cards - Hugging Face, accesso eseguito il giorno settembre 19, 2025, https://huggingface.co/docs/hub/datasets-cards
6. MedGemma model card | Health AI Developer Foundations, accesso eseguito il giorno settembre 19, 2025, https://developers.google.com/health-ai-developer-foundations/medgemma/model-card