

Percorsi di Oncologia di Precisione:

Appropriatezza diagnostica e
Molecular Tumor Board

Oncologia molecolare: ruolo dei Registri Tumori

Manuel ZORZI

Registro Tumori del Veneto
Azienda Zero - Padova



*Il sottoscritto **Manuel Zorzi**
in qualità di relatore all'evento*

Percorsi di Oncologia di Precisione

Milano, 30 gennaio 2026

ai sensi dell'art. 76, comma 4 dell'Accordo Stato Regioni del 2 febbraio 2017 e del paragrafo 4.5. del Manuale nazionale di accreditamento per l'erogazione di eventi ECM

dichiara che negli ultimi due anni non ha avuto rapporti con soggetti portatori di interessi commerciali in ambito sanitario.

Cos'è un REGISTRO TUMORI

- Struttura per la raccolta sistematica di informazioni sulle **nuove diagnosi di tumore** nei **cittadini residenti in una certa area**.
- Lo scopo è **identificare tutti i tumori** che vengono diagnosticati in un certo periodo nella popolazione residente, definendo con precisione **quante persone** si sono ammalate di tumore, che **tipo di tumore** hanno e che **decorso** hanno avuto.

Valore aggiunto dei dati di popolazione

- Quantificazione dei fenomeni ai fini di **programmazione sanitaria**
- Le casistiche ospedaliere non garantiscono la **rappresentatività dei fenomeni**: selezione di pazienti in funzione del livello del centro di ricerca, perdita selettiva di sottogruppi di pazienti in genere più problematici, a causa di caratteristiche individuali (età avanzata, comorbidity, ...) o di problemi di accessibilità (classe socio-economica, gruppi hard-to-reach, ...)
- Real world data di popolazione: indispensabili per valutare l'effettivo **rispetto di protocolli, linee guida, PDTA** da parte della totalità dei pazienti e gli esiti di salute

... e inoltre...

- Disponibilità di fonti di dati su tutta la popolazione oncologica.
Quindi possibilità di contribuire al recupero di **informazioni sul follow up di casistiche ospedaliere** (terapie, esiti, ecc).

Fonti di dati

- Poiché **il percorso diagnostico e terapeutico** dei pazienti, anche affetti dallo stesso tipo di tumore, **può essere molto diverso**, per ottenere informazioni complete, i RT utilizzano diverse fonti informative.
- Quelle fondamentali sono le **schede di dimissione ospedaliera (SDO), i referti di anatomia patologica (AP) e i certificati di morte.**
- Come **fonti accessorie** sono inoltre utilizzati gli archivi di radiodiagnostica, le prestazioni specialistiche, le esenzioni per patologia e i flussi relativi alla farmaceutica.

Veneto: FLUSSO ANAPAT (anno 2016)

Il Flusso Anapat è un flusso di **referti** di Anatomia Patologica **la cui diagnosi è stata **codificata** come neoplasia**

Le informazioni sono divise in tre archivi:

1. Dati **anagrafici**
2. Dati **sanitari codificati**
3. Dati **sanitari in chiaro**

I flussi informativi: esperienza del Friuli Venezia Giulia



IRCCS CRO
Istituto di ricovero e cura a carattere
scientifico "Centro di Riferimento
Oncologico" di Aviano



REGIONE AUTONOMA FRIULI VENEZIA GIULIA

- Strutturazione di un **data warehouse epidemiologico regionale** organizzato secondo una architettura orientata alla rilevazione degli eventi sanitari della popolazione.
- Insiel ha scelto un'unica piattaforma tecnologica (SAS) per:
 - la realizzazione del sistema;
 - lo sviluppo dell'algoritmo che calcola l'incidenza;
 - l'interfaccia risoluzione manuale dei casi;
 - le attività di analisi a fini statistici e di programmazione per l'intero sistema sanitario regionale.

Nella piattaforma informativa epidemiologica:

- Anatomia patologica (referti dal 1982)
- Dimissione ospedaliera (lettere dal 2008)
- Cartella oncologica (testi dal 2005)
- Prestazioni cliniche (referti dal 2000)

Courtesy L. Dal Maso

Metodi

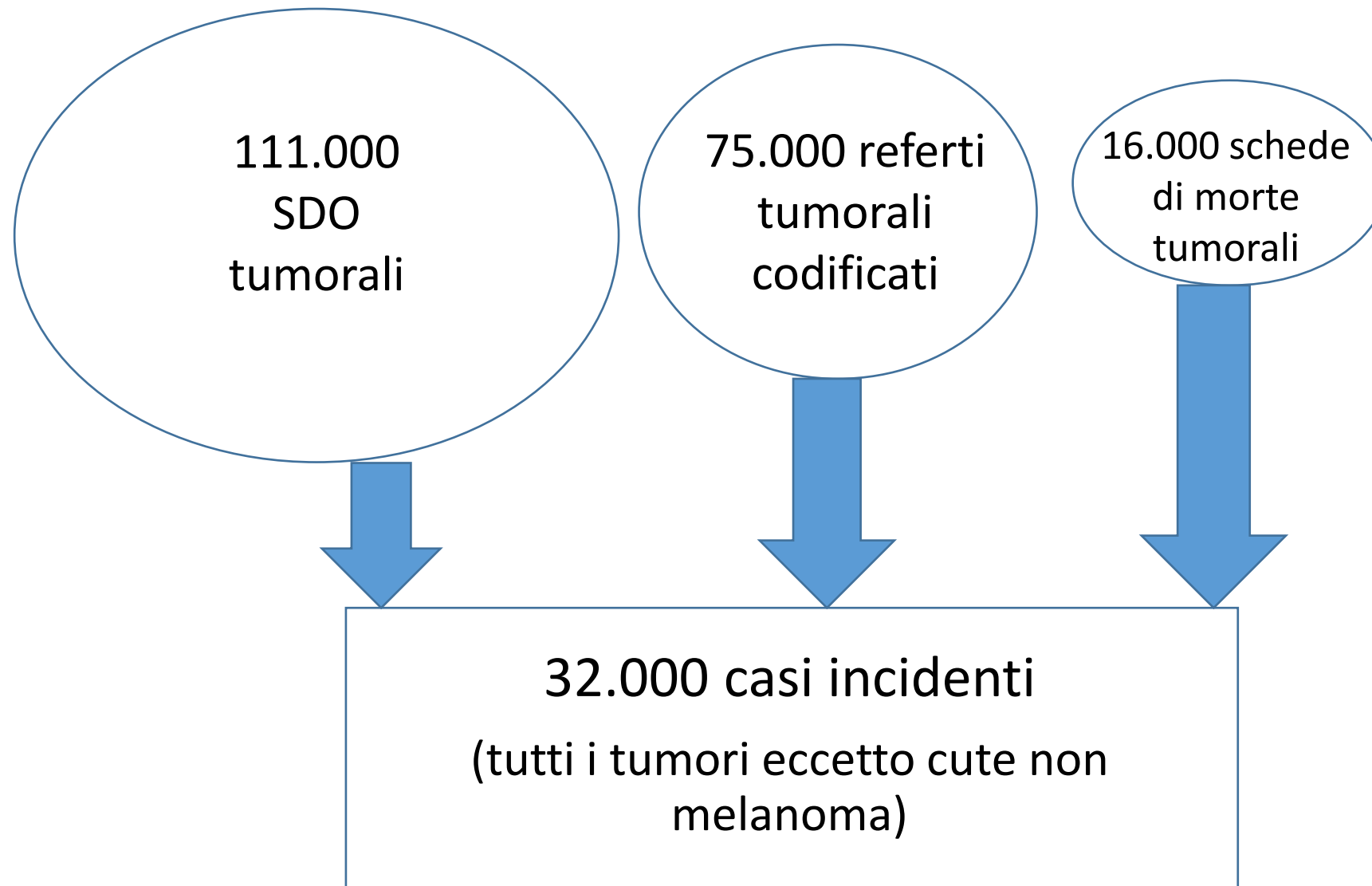
- Il RTV utilizza un sistema di registrazione dell'incidenza basato sui **dati codificati e informatizzati relativi alle dimissioni ospedaliere, i referti di anatomia patologica e i certificati di morte.**
- Tali fonti diagnostiche vengono inizialmente incrociate con **l'anagrafe sanitaria regionale**, per selezionare i soggetti residenti in Veneto.
- Dagli archivi così selezionati vengono quindi **eliminati i casi prevalenti**: in sede di prima registrazione, quando tutte le date riferite dalle fonti sono anteriori al periodo d'incidenza; in sede di aggiornamento, quando le nuove segnalazioni riferiscono diagnosi eguali o compatibili con i tumori già registrati.

Sistema di registrazione dell'incidenza delle neoplasie

Registro Tumori del Veneto



Anno 2021 - Veneto



I Registri Tumori raccolgono **informazioni limitate** (sesso, età, **topografia e morfologia, data di incidenza, stato in vita**)

Obiettivi:

- Descrivere l'epidemiologia di base dei tumori
- Monitorare i trend temporali e spaziali dei tumori
- Mostrare gli effetti degli interventi di politica sanitaria: es. pratiche di prevenzione



OPEN ACCESS

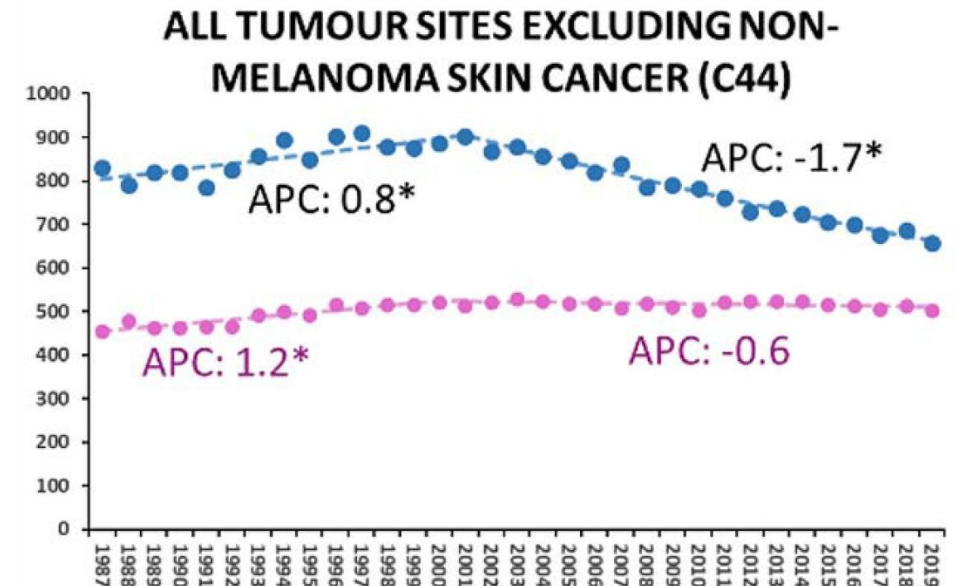
EDITED BY
Ingmar Schäfer,
University Medical Center Hamburg-Eppendorf,
Germany

REVIEWED BY
Annalisa Quattrocchi,
University of Nicosia, Cyprus
Yufei Liu,
Shenzhen Second People's Hospital, China

*CORRESPONDENCE

Thirty-two-year trends of cancer incidence by sex and cancer site in the Veneto Region from 1987 to 2019

Alessandra Buja^{1*}, Giuseppe De Luca¹, Manuel Zorzi²,
Emanuela Bovo², Simone Mocellin^{3,4}, Chiara Trevisiol³,
Vincenzo Bronte⁵, Stefano Guzzinati² and Massimo Rugge^{2,6}



ONCOLOGIA MOLECOLARE

Il **problema limitante** è la mancata disponibilità di dati sulle analisi di profilazione molecolare

- **Mancano flussi**, o anche sistemi di archiviazione dei referti da parte dei laboratori (FSE non accessibile!)
- Dove disponibili i referti (es. all'interno di cartelle cliniche, ...), i **dati non sono strutturati**, pertanto necessitano di una consultazione da parte di un operatore

I flussi informativi: esperienza del Friuli Venezia Giulia



IRCCS CRO
Istituto di ricovero e cura a carattere
scientifico "Centro di Riferimento
Oncologico" di Aviano

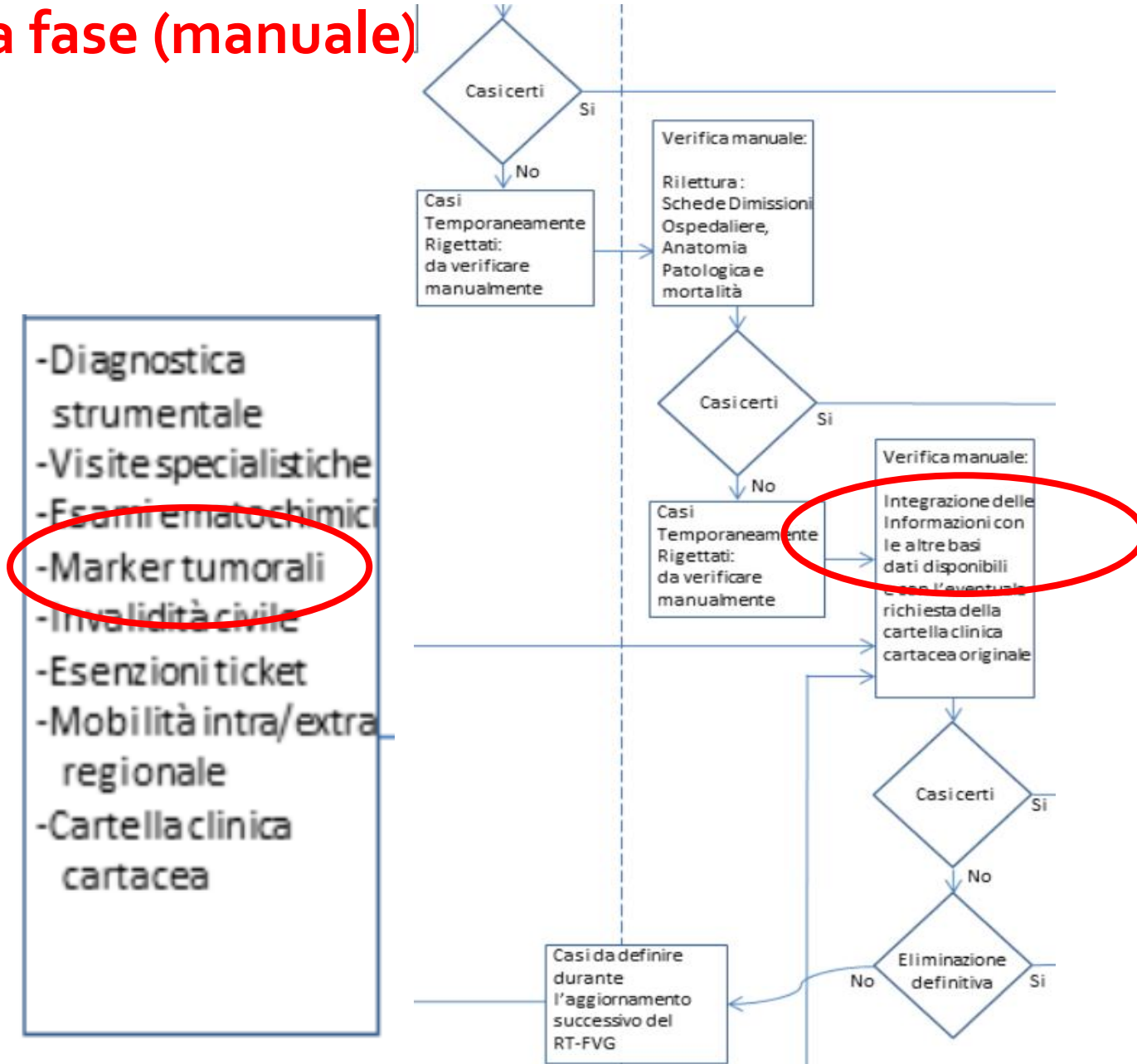


- Strutturazione di un **data warehouse epidemiologico regionale** organizzato secondo una architettura orientata alla rilevazione degli eventi sanitari della popolazione.
- Insiel ha scelto un'unica piattaforma tecnologica (SAS) per:
 - la realizzazione del sistema;
 - lo sviluppo dell'algoritmo che calcola l'incidenza;
 - l'interfaccia risoluzione manuale dei casi;
 - le attività di analisi a fini statistici e di programmazione per l'intero sistema sanitario regionale.

-Schede Dimissioni
Ospedaliere
-Anatomia
Patologica
-Decessi
-RT-FVG
-Cartella clinica
oncologica

-Diagnostica
strumentale
-Visite specialistiche
-Esami ematochimici
-Marker tumorali
-Invalidità civile
-Esenzioni ticket
-Mobilità intra/extra
regionale
-Cartella clinica
cartacea

Seconda fase (manuale)



Dati raccolti dal Registro Tumori del Lussemburgo



Data type	Examples of data items collected
Record Identification	- Registry Identification Number
Demographic data	- Age at diagnosis - Gender - Country of birth - Last known address
Death data	- Date of death - Cause of death - Autopsy
Tumor data	- Date of incidence - Topography (*ICD-O-3) - Morphology (ICD-O-3) - Basis of diagnosis - Clinical stages - Pathological stages - Metastases at time of diagnosis - Biopsy related data - Cytology related data - Histological prognostic factors - Tissue tumor markers and molecular alterations
Clinical data	- Circumstances of discovery - Comorbidity - Performance score (*ECOG)
Therapeutic management data	- Initial treatment - Surgery - Chemotherapy - Hormone therapy - Radiotherapy - Targeted therapy - Other treatments

*ICD-O-3, International classification of diseases for oncology (ICD-O) – 3rd edition,

*ECOG, Eastern Cooperative Oncology Group.

Dati raccolti dai Registri Tumori Italiani



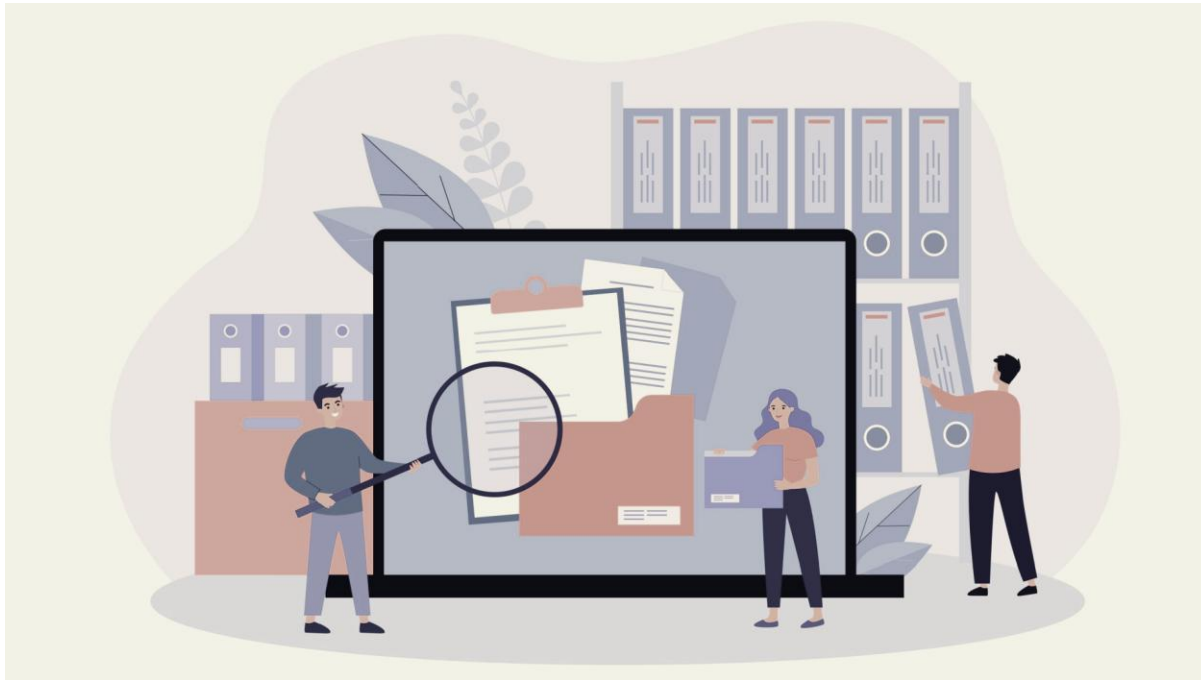
Data type	Examples of data items collected
Record Identification	- Registry Identification Number
Demographic data	- Age at diagnosis - Gender - Country of birth - Last known address
Death data	- Date of death - Cause of death - Autopsy
Tumor data	- Date of incidence - Topography (*ICD-O-3) - Morphology (ICD-O-3) - Basis of diagnosis - Clinical stages - Pathological stages - Metastases - Biopsy - Cytology - Histology - Tumor markers and molecular alterations
Clinical data	- Circumstances of diagnosis - Comorbidity - Performance score (ECOG)
Therapeutic management data	- Initial treatment - Subsequent treatments - Chemotherapy - Hormonal therapy - Radiation therapy - Other treatments

*ICD-O-3, International classification of diseases for oncology (ICD-O) – 3rd edition,

*ECOG, Eastern Cooperative Oncology Group.

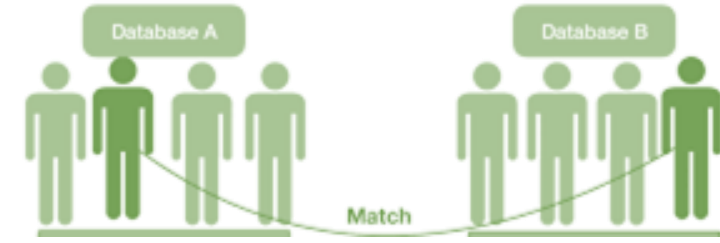
REGISTRAZIONE AD ALTA RISOLUZIONE

Fonti dei dati che il data manager consulta



1. Anatomia patologica
2. Referti radiologici
3. Archivi ospedalieri
4. Cartelle cliniche

DATABASE COMPLETO



RECORD LINKAGE

Registro ad alta risoluzione

- Anatomia patologica
- Referti radiologici
- Archivi ospedalieri
- Cartelle cliniche

Flussi informativi sanitari

- Schede dimissioni ospedaliere
- Specialistica ambulatoriale
- Farmaci ad alto costo (File F)
- Farmaceutica territoriale
- Pronto soccorso
- Dispositivi Medici
- Hospice
- ADI

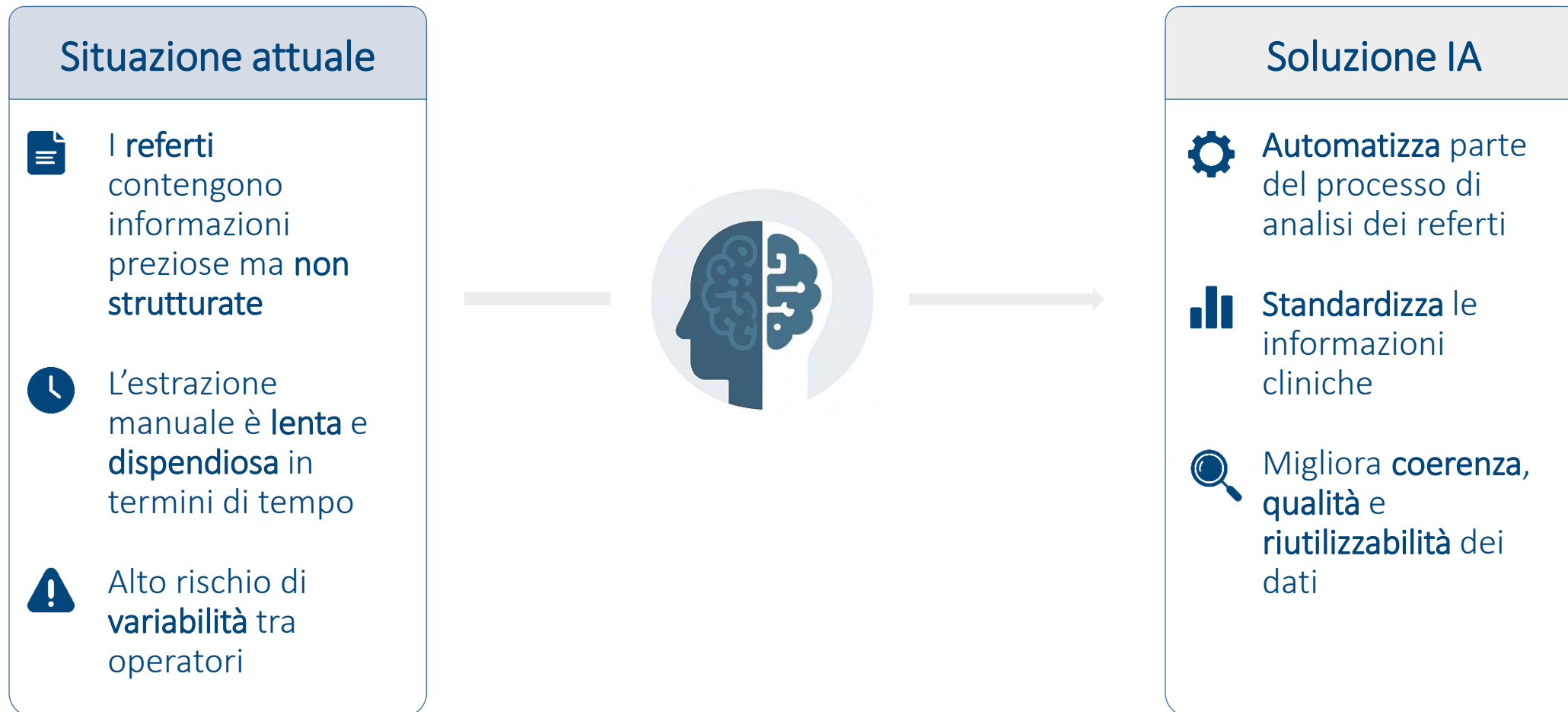
Gli archivi ad Alta Risoluzione possono sono utilizzati per diversi scopi

1. Supporto per **studi epidemiologici**
2. Supporto per **studi clinici**
3. Supporto per studi volti a **valutare i costi** del percorso di cura
4. Supporto per studi volti a **valutare la qualità** del percorso diagnostico terapeutico

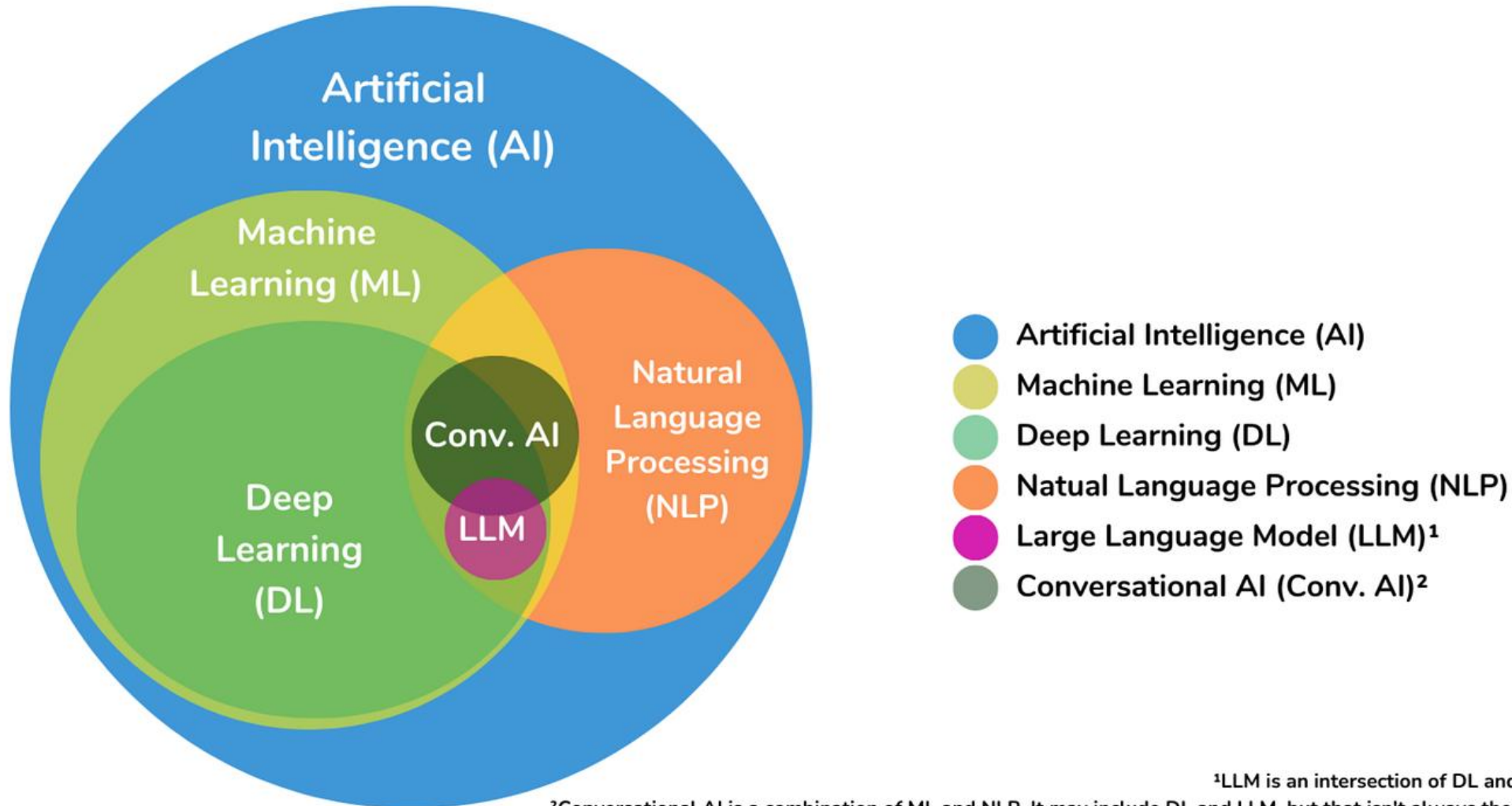


Sviluppo di metodologie di intelligenza artificiale per estrarre dati da testi liberi

Perché utilizzare l'Intelligenza Artificiale (IA) in ambito clinico



Dall'AI agli LLM: l'evoluzione dell'intelligenza artificiale nel linguaggio



Il processo di Text Mining (TM): dal testo libero al dato strutturato

Il Text Mining permette di convertire conoscenza clinica non strutturata in dati strutturati facilmente analizzabili



Pulizia del testo

Rimozione di simboli, stopwords, abbreviazioni



Estrazione

Identificazione di parole chiave e concetti



Classificazione

Assegnazione a categorie cliniche o tematiche



Analisi

Creazione di dati strutturati per analisi statistiche

Obiettivo e metodologia adottata

Estrazione dei biomarcatori dai referti di anatomia patologica



Contesto e obiettivo

Referti **testuali** di anatomia patologica dei **casi incidenti 2017-2020**

Estrazione dei **biomarcatori** del **tumore della mammella femminile** (ER, PgR, HER2, Ki-67) dai referti di anatomia patologica



Metodologia adottata

Pulizia e **normalizzazione** del testo

Estrazione di sottostringhe mediante **parole chiave**

Applicazione del **Text Mining**

Applicazione di un algoritmo di **Machine Learning**



Validazione dei risultati

Confronto tra **valori predetti** e **gold standard**

Calcolo delle metriche di **accuratezza** (F1-score pesato)

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Esempio di un campo diagnosi nei referti di AP

Alcuni minuti frustoli di parenchima mammario, tre dei quali sede di carcinoma infiltrante di tipo non speciale (duttale, NAS), il cui grado non è valutabile a causa dell'esiguità del reperto. La lesione misura cm 0,35 nel frustolo maggiore. Componente intraduttale non rappresentata. Categoria diagnostica: B5b (Lesioni maligne) (Sec. European guidelines for quality assurance in breast cancer screening and diagnosis- IV Edizione) Fenotipo: - Proteina recettore estrogenico presente nel 50% degli elementi neoplastici. - Proteina recettore progestinico presente nel 30% degli elementi neoplastici. - Marcatore di proliferazione Ki67 (MIB1) positivo nel 40% degli elementi neoplastici. - HER2 Negativo: Score 0. Procedura ER, PGR, KI67: Smascheramento antigenico con PTlink a 95C° per 10 min (EnVision FLEX) Immunocolorazione con sistema automatizzato DAKO Autostainer Recettore estrogenico: clone Sp1 (Dako) Recettore progestinico: clone 636 (Dako) Proteina Ki67: MIB1 (Dako) HER2: Smascheramento antigenico con bagno termostato preriscaldato a 98C° per 40 min Immunocolorazione con sistema automatizzato DAKO Autostainer e con HERCEPTEST DAKO Lettura effettuata secondo le raccomandazioni ASCO/CAP 2013 per la determinazione dello stato di HER2 nel carcinoma della mammella Positivo 3+ = positività di membrana circonferenziale completa, uniforme ed intensa in più del 10% degli elementi Incerto 2+ = positività di membrana circonferenziale completa intensa in meno del 10% degli elementi o positività di me

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Esempio di un campo diagnosi nei referti di AP

Alcuni minuti frustoli di parenchima mammario, tre dei quali sede di carcinoma infiltrante di tipo non speciale (duttale, NAS), il cui grado non è valutabile a causa dell'esiguità del reperto. La lesione misura cm 0,35 nel frustolo maggiore. Componente intraduttale non rappresentata. Categoria diagnostica: B5b (Lesioni maligne) (Sec. European guidelines for quality assurance in breast cancer screening and diagnosis- IV Edizione) Fenotipo: - **Proteina recettore estrogenico presente nel 50% degli elementi neoplastici.** - **Proteina recettore progestinico presente nel 30% degli elementi neoplastici.** - **Marcatore di proliferazione Ki67 (MIB1) positivo nel 40% degli elementi neoplastici.** - **HER2 Negativo: Score 0.** Procedura ER, PGR, KI67: Smascheramento antigenico con PTlink a 95C° per 10 min (EnVision FLEX) Immunocolorazione con sistema automatizzato DAKO Autostainer Recettore estrogenico: clone Sp1 (Dako) Recettore progestinico: clone 636 (Dako) Proteina Ki67: MIB1 (Dako) HER2: Smascheramento antigenico con bagno termostato preriscaldato a 98C° per 40 min Immunocolorazione con sistema automatizzato DAKO Autostainer e con HERCEPTEST DAKO Lettura effettuata secondo le raccomandazioni ASCO/CAP 2013 per la determinazione dello stato di HER2 nel carcinoma della mammella Positivo 3+ = positività di membrana circonferenziale completa, uniforme ed intensa in più del 10% degli elementi Incerto 2+ = positività di membrana circonferenziale completa intensa in meno del 10% degli elementi o positività di me

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Obiettivo e metodologia adottata

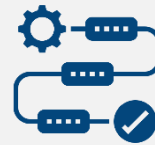
Estrazione dei biomarcatori dai referti di anatomia patologica



Contesto e obiettivo

Referti **testuali** di anatomia patologica dei casi incidenti 2017-2020

Estrazione dei biomarcatori del tumore della mammella femminile (ER, PgR, HER2, Ki-67) dai referti di anatomia patologica



Metodologia adottata

Pulizia e normalizzazione del testo

Estrazione di sottostringhe mediante **parole chiave**

Applicazione del **Text Mining**

Applicazione di un algoritmo di **Machine Learning**



Validazione dei risultati

Confronto tra valori predetti e gold standard

Calcolo delle metriche di accuratezza (F1-score pesato)

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Obiettivo e metodologia adottata

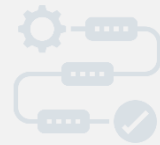
Estrazione dei biomarcatori dai referti di anatomia patologica



Contesto e obiettivo

Referti **testuali** di anatomia patologica dei casi incidenti 2017-2020

Estrazione dei biomarcatori del tumore della mammella femminile (ER, PgR, HER2, Ki-67) dai referti di anatomia patologica



Metodologia adottata

Pulizia e normalizzazione del testo

Estrazione di sottostringhe mediante **parole chiave**

Applicazione del **Text Mining**

Applicazione di un algoritmo di **Machine Learning**



Validazione dei risultati

Confronto tra **valori predetti** e **gold standard**

Calcolo delle metriche di **accuratezza** (F1-score pesato)

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Risultati principali

Metriche di validazione sul test set (20%)

Metriche	Biomarcatore			
	ER	PgR	Ki-67	HER2
Numero di pazienti	3.613	3.592	3.540	3.574
F1 pesato (valori esatti)	89,3%	89,1%	87,2%	91,7%
F1 pesato (valori ricodificati*)	99,6%	98,1%	96,7%	95,4%

* Si intende <10%/≥10% per ER, <20%/≥20% per PGR e Ki-67, e negativo/dubbio/positivo/mancante per Her2

CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Risultati principali

Metriche di validazione sui casi 2021





Metriche	Biomarcatore			
	ER	PgR	Ki-67	HER2
Numero di pazienti	935	867	827	749
F1 pesato (valori esatti)	76,6%	57,8%	66,8%	76,0%
F1 pesato (valori ricodificati*)	98,0%	88,7%	91,4%	87,9%

* Si intende <10%/≥10% per ER, <20%/≥20% per PGR e Ki-67, e negativo/dubbio/positivo/mancante per Her2





CASO STUDIO: TUMORE DELLA MAMMELLA FEMMINILE

Conclusioni

Punti di forza

-  **Automazione del processo:** riduzione del carico di lavoro manuale e aggiornamento continuo dei dati
-  Applicazione innovativa in **lingua italiana**, ambito ancora poco esplorato nella letteratura scientifica
-  **Utilizzo di R:** soluzione open-source, intuitiva e integrabile con altri algoritmi di classificazione
-  **Metodo generalizzabile:** approccio estendibile ad altri tipi di tumori o variabili cliniche

Limiti

-  **Copertura parziale del campione:** referti provenienti solo da 7 su 21 servizi di Anatomia Patologica del Veneto
-  **Differenze di formato dei dati:** i registri tumori includono anche documenti non testuali (PDF, immagini), non gestibili dal modello
-  **Un solo modello ML:** mancano confronti con altri algoritmi (Random Forest, Gradient Boosting)
-  **Limiti di accuratezza legati alla dimensione del dataset:** i modelli migliorano con l'aumento dei casi di addestramento

ALGORITMO DI TEXT MINING RULE-BASED

Nel Text Mining rule-based l'analisi dei testi si basa su regole esplicite definite da esperti



Pulizia del testo

Rimozione di simboli,
stopwords,
abbreviazioni



Creazione di dizionari

Elenco di parole chiave
o frasi rilevanti



Definizione di regole

Definite manualmente
al fine di estrarre
correttamente le
informazioni dal testo



Estrazione

Applicazione delle
regole ai testi



Output strutturato

Risultati pronti per
analisi o integrazione
in database

CASO STUDIO: GLEASON – TUMORE DELLA PROSTATA

Risultati principali

PDTA PROSTATA – pazienti con referti AP nel 2021-2022

Gleason Score	N
Gleason score 4 (2+2)	2
Gleason score 5 (2+3)	2
Gleason score 5 (3+2)	7
Gleason score 6 (3+3)	3.282
Gleason score 7 (3+4)	1.385
Gleason score 7 (4+3)	943
Gleason score 8 (3+5)	29
Gleason score 8 (4+4)	746
Gleason score 8 (5+3)	32
Gleason score 9 (4+5)	175
Gleason score 9 (5+4)	73
Gleason score 10 (5+5)	31
Gleason score non applicabile	28
Gleason score non presente nella diagnosi	591
Totale	7.326



CASO STUDIO: GRADO TUMORALE – TUMORI SNC

Risultati principali

Pazienti incidenti per tumore SNC nel 2016-2020

		Grado 2	Grado 3	Grado 4
1.636 casi	1.056 (64,6%)			1.056 (100,0%)
	Glioblastoma IDH-wildtype e IDH-mutato	-	-	
	154 (9,4%)	62 (40,3%)	92 (59,7%)	-
	Astrocitoma grado 2-3			
	298 (18,2%)	272 (91,3%)	26 (8,7%)	-
	Meningioma grado 2-3			
	74 (4,5%)	36 (48,7%)	38 (51,3%)	-
	Oligodendroglioma grado 2-3			
	39 (2,4%)	34 (87,2%)	5 (12,8%)	-
	Ependimoma grado 2-3			
	15 (0,9%)	-	-	15 (100,0%)
	Tumore embrionale del SNC (medulloblastoma)			
		404 (24,7%)	161 (9,8%)	1.071 (65,5%)

CASO STUDIO: STADIO T & N – TUMORE DELLA VESCICA

Risultati principali

PDTA VESCICA - Pazienti incidenti per tumore invasivo della vescica nel 2021 (ICD10 C67)

Stadio T	M		F		Totale	
	N	%	N	%	N	%
T0	2	0,7	0	0,0	2	0,6
Ta	51	18,8	15	19,0	66	18,8
Tis	2	0,7	1	1,3	3	0,9
T1	81	29,8	13	16,5	94	26,8
T2	32	11,8	12	15,2	44	12,5
T2a	18	6,6	2	2,5	20	5,7
T2b	9	3,3	4	5,1	13	3,7
T3	7	2,6	3	3,8	10	2,8
T3a	18	6,6	12	15,2	30	8,5
T3b	18	6,6	13	16,5	31	8,8
T4	2	0,7	0	0,0	2	0,6
T4a	32	11,8	4	5,1	36	10,3
Totale	272	100,0	79	100,0	351	100,0

Stadio N	M		F		Totale	
	N	%	N	%	N	%
Nx	12	10,2	4	12,9	16	10,7
N0	64	54,2	15	48,4	79	53,0
N1	18	15,3	5	16,1	23	15,4
N2	23	19,5	6	19,4	29	19,5
N3	1	0,8	1	3,2	2	1,3
Totale	118	100,0	31	100,0	149	100,0

Text Mining: approccio rule-based e approccio Machine Learning

Text Mining & modelli ML

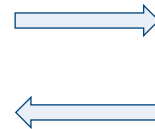


Algoritmi addestrati su informazioni registrate a priori (**gold standard**)

Maggiore **adattabilità** e **generalizzazione**

Efficace su **testi complessi**

Richiede **molti dati** e **risorse computazionali** elevate



Text Mining rule-based



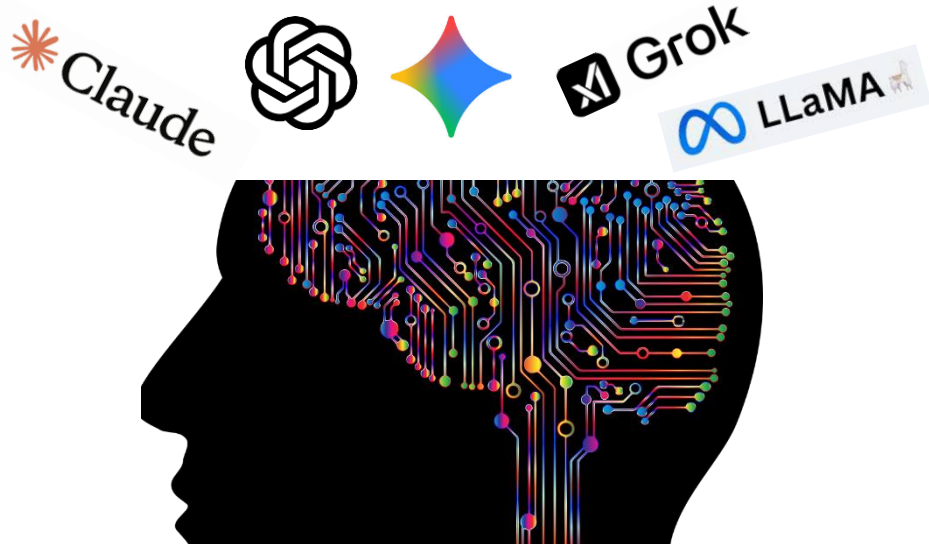
Parole chiave e regole linguistiche definite **manualmente**

Alta **trasparenza** e **controllo** (non ci sono «scatole nere»)

Efficace su **testi strutturati**

Richiede **manutenzione** continua

LARGE LANGUAGE MODEL (LLM)



Gli LLM sono modelli di AI basati sull'apprendimento profondo (*Deep Learning*).

- 👉 Possiamo immaginarlo come una rete di “neuroni artificiali”, ispirata al cervello umano.
- 👉 La rete viene “addestrata” con enormi quantità di testi e durante questo processo impara da sola a riconoscere schemi, collegamenti e significati nel linguaggio.

Come ogni modello linguistico ...

l'obiettivo di base degli LLM è **predire il token successivo in una sequenza di parole**, basandosi sui pattern linguistici appresi durante l'addestramento.

Il cielo è...**blu**

Perché «Large»?

- Gli LLM sono addestrati su **enormi quantità di testo** (per riconoscere ampie sfumature linguistiche, stili e contesti)
- Si basano su una rete neurale profonda («cervello artificiale») con un **numero di parametri** (nodi e connessioni) che superano il trilione)

CASO STUDIO 2: Applicazione degli LLM



OBIETTIVO

Sviluppare e validare un sistema automatizzato per l'estrazione e la classificazione della **stadiazione del tumore alla mammella femminile**



POPOLAZIONE

Casi incidenti di tumore alla mammella tra le donne residenti nella Regione Veneto (2017-2021)



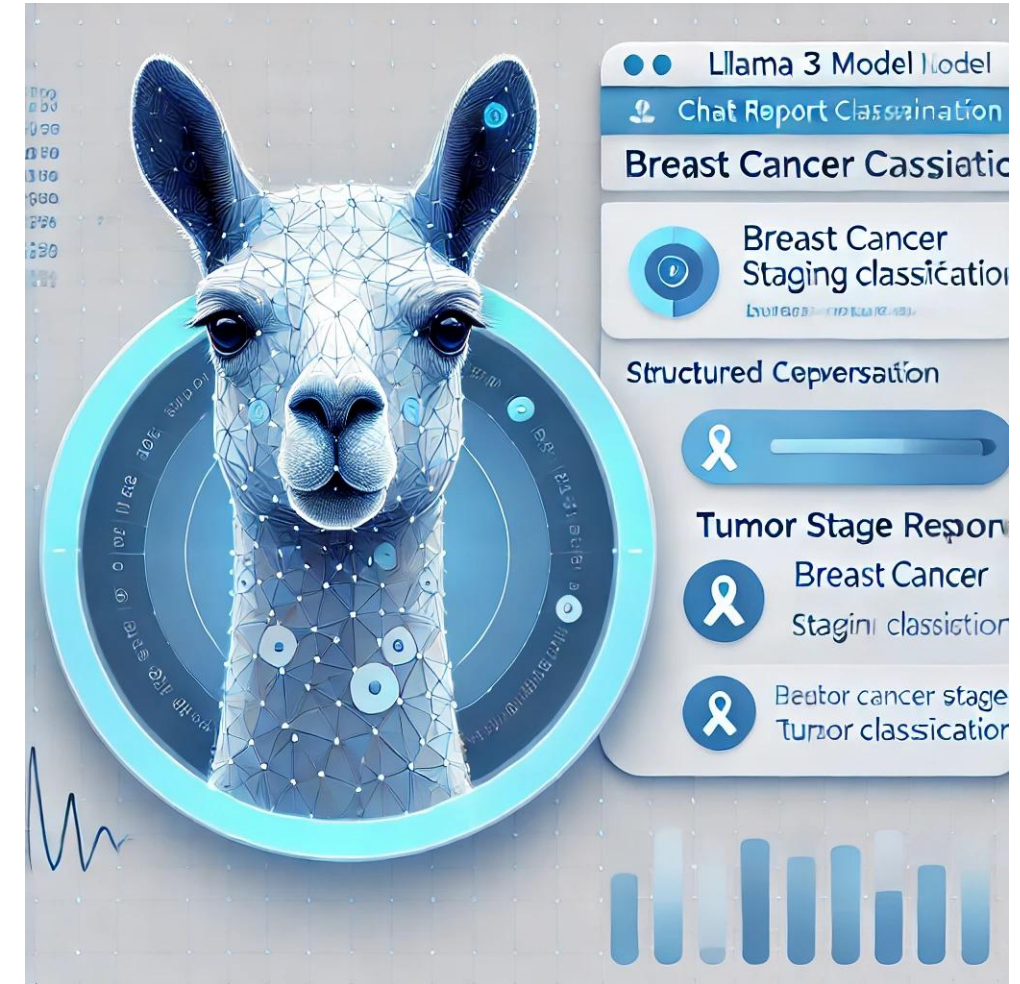
DATA SOURCE

- Referti di anatomia patologica
- Referti di radiodiagnostica

- Il modello ha ottenuto una classificazione quasi perfetta in tutte le categorie.

	Accuracy	Weighted precision	Weighted recall	Weighted F1-score
T-category				
Training	1.0000	1.0000	1.0000	1.0000
Fine-tuning	1.0000	1.0000	1.0000	1.0000
Validation	1.0000	1.0000	1.0000	1.0000
N-category				
Training	0.9997	0.9997	0.9997	0.9997
Fine-tuning	1.0000	1.0000	1.0000	1.0000
Validation	1.0000	1.0000	1.0000	1.0000
M-category				
Training	0.9996	0.9996	0.9996	0.9996
Fine-tuning	1.0000	1.0000	1.0000	1.0000
Validation	1.0000	1.0000	1.0000	1.0000

- Durante la **fase di training**, si sono verificati un errore di classificazione nella classificazione T e uno nella classificazione M, dovuti ad errore umano nelle annotazioni dei dati di origine.
- Durante le fasi di **fine-tuning e validation**, non si sono verificati errori, ad indicare che il modello ha appreso e applicato efficacemente i criteri di classificazione TNM.



Cosa ci portiamo a casa? TM tradizionale vs LLM



Rappresentazione del Linguaggio

Vettori Numerici Statici (e.g., TF-IDF).
I testi vengono convertiti in rappresentazioni statiche del linguaggio e analizzati come insiemi di parole indipendenti, senza tener conto di significato e contesto.

Embedding Contestuali.

Vettori densi che catturano relazioni semantiche e sintattiche profonde, l'ordine delle parole, sfumature del linguaggio e contesto

Ambiguità Linguistica (Esempio)

Non gestita.
La parola "banca" in "banca del sangue" e "banca di credito" viene rappresentata allo stesso modo, perdendo il significato.

Gestita.

Le rappresentazioni (vettori) della parola "banca" cambiano in base al contesto della frase

Applicabilità

Ideale per **estrarre dati strutturati** e **classificare** in contesti semplici, dove l'ambiguità è minima.

Essenziale per **comprendere, sintetizzare o generare conoscenza** da testi complessi e fortemente contestuali.

L'utilizzo dell'IA per il text mining risolve i limiti attuali?



Mancano flussi, o anche sistemi di archiviazione dei referti da parte dei laboratori (FSE non accessibile!)



Dove disponibili i referti (es. all'interno di cartelle cliniche, ...), i **dati non sono strutturati**, pertanto necessitano di una consultazione da parte di un operatore

L'utilizzo dell'IA per il text mining introduce nuove criticità?

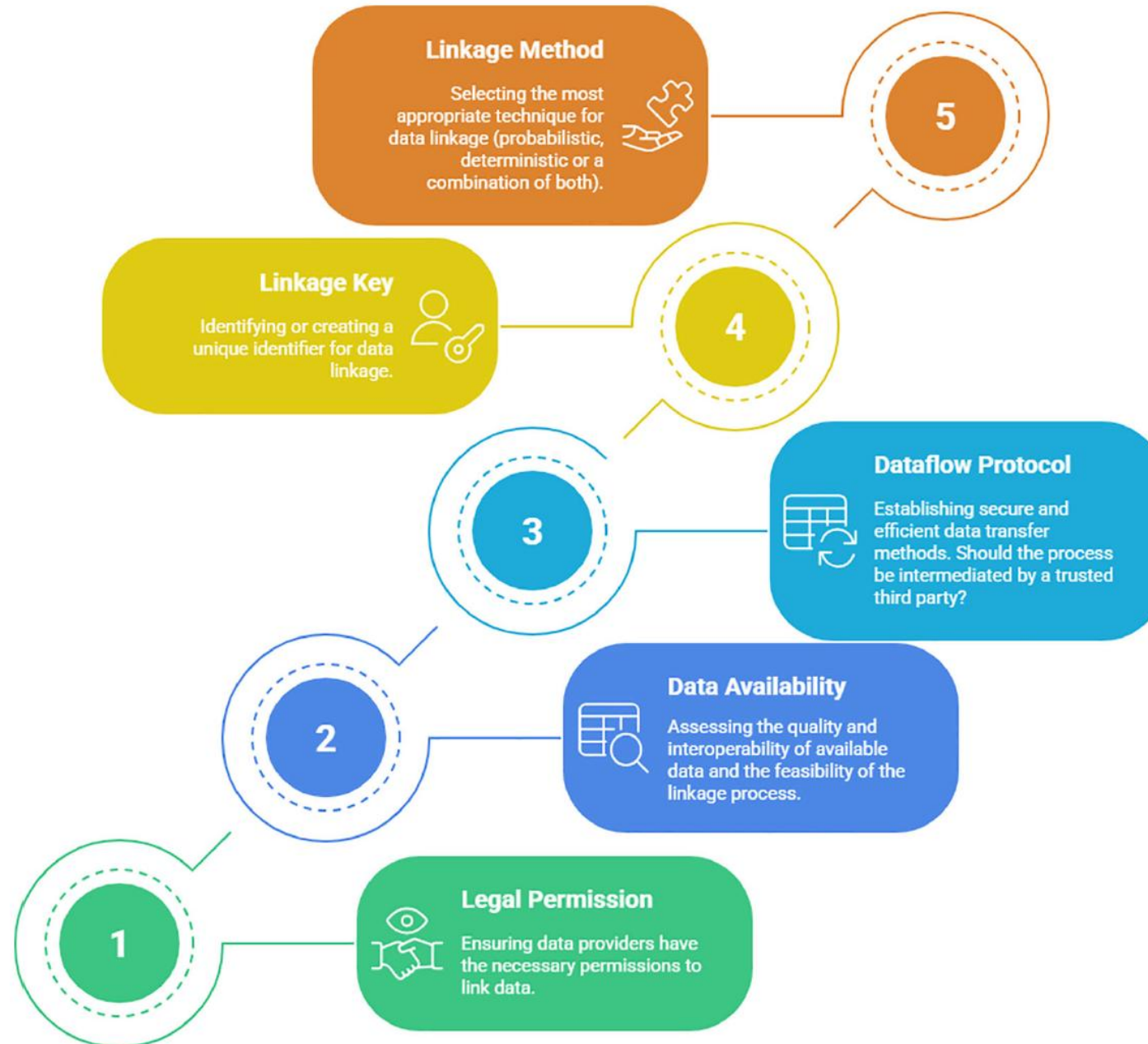


Risorse e knowhow per sviluppare gli algoritmi



Performance degli algoritmi, problemi di validazione, ecc

Key domains and related questions to be considered, prior to establishing data linkage



DECRETO 1° agosto 2023.

Registro nazionale tumori.

Tipi di dati personali trattati

2. Per il perseguimento delle finalità di cui all'art. 3, vengono trattati dati personali relativi alla salute, riferiti a casi diagnosticati di neoplasia, di cui all'art. 1, comma 1, lettera *b*), trasmessi dai Centri di riferimento regionali sulla base dei dati individuali relativi a:

a) diagnosi di ammissione e dimissione, relative a ricoveri e a prestazioni diagnostico-terapeutiche;

b) modalità di dimissione relative ai ricoveri;

c) anamnesi;

d) interventi chirurgici e procedure diagnostiche e terapeutiche, ivi compresi gli *screening* oncologici;

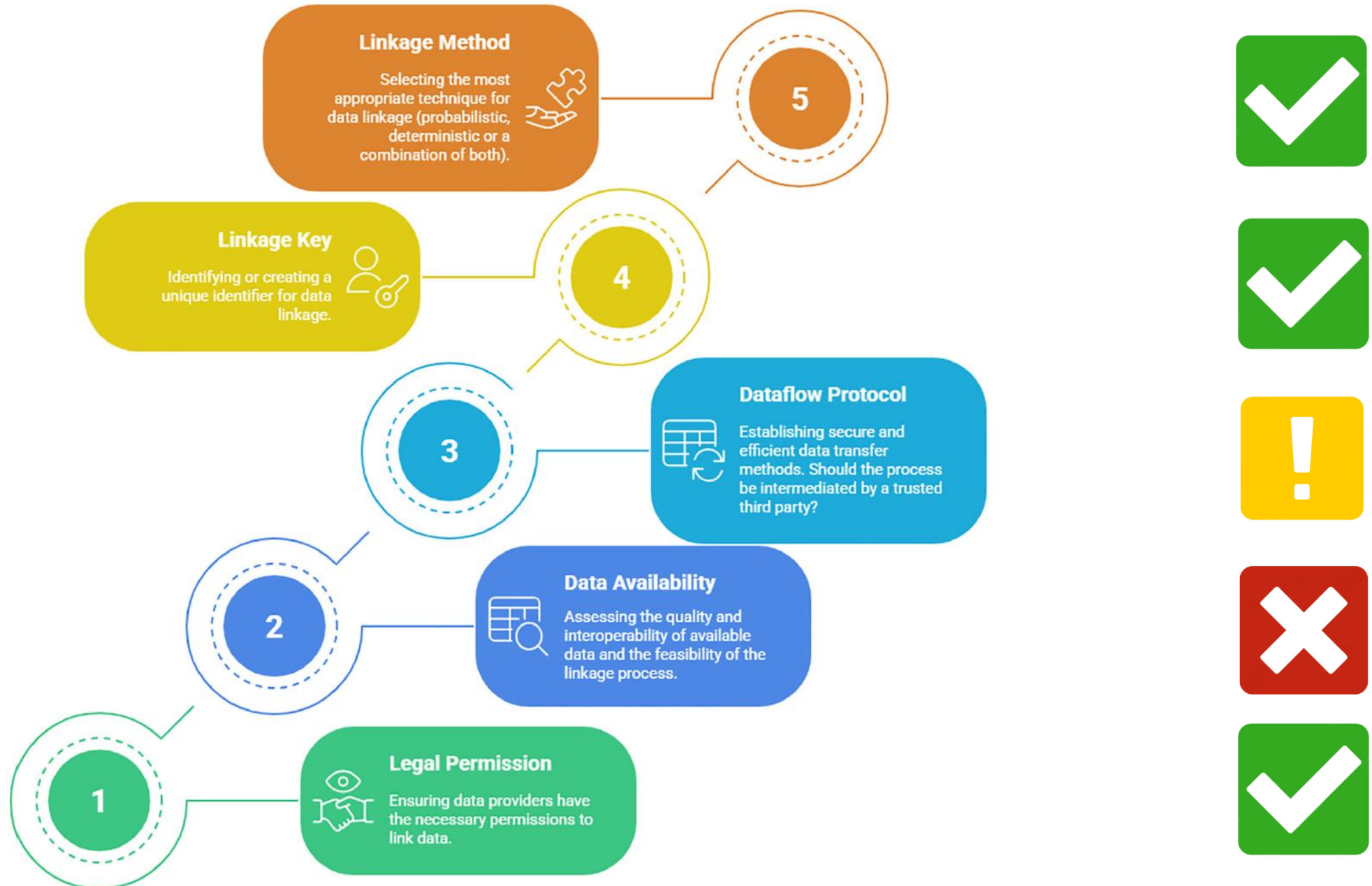
e) indagini e analisi cliniche e strumentali e i trattamenti eseguiti;

f) referti anatomo-patologici con indicazione della sede, morfologia, grado di differenziazione e comportamento biologico, comprese le indagini biomolecolari atte a definire la prognosi della neoplasia;

g) tecniche di definizione diagnostica;

h) data e causa di morte e condizioni morbose rilevanti per il decesso.

Key domains and related questions to be considered, prior to establishing data linkage



AGENDA

- Strutturare gli archivi dei referti di biologia molecolare
- Strutturare i dati all'interno degli archivi
- Strutturare la trasmissione degli archivi ai Registri Tumori di riferimento

Esempio: proposta di obiettivi 2026 dei Direttori Generali delle Aziende Sanitarie

Linea_strategica	Q - Crescita dei livelli di qualità dell'assistenza
Ambito LEA	S - Processi di supporto
Obiettivo	Miglioramento efficienza dei processi di supporto
Descrizione Indicatore	Percentuale di referti di Anatomia Patologica con diagnosi tumorale che valorizzano i campi pT e/o pN nel flusso ANAPAT
Metodo calcolo	Referti nel flusso ANAPAT. Numeratore: referti con campo pT e/o pN valorizzato. Denominatore: totale referti selezionati. Criteri di selezione: Morfologia = tumore maligno AND Procedura = secondo un elenco da allegare al Vademecum
Criterio soddisfazione soglia	> 80%
Scadenza	31/08/2026 (referti del primo semestre 2026) e 28/02/2027 (referti dell'intero 2026)

Grazie per l'attenzione