Trappole e Rischi Dell'Intelligenza Artificiale In Medicina.

Prof. Francesco Macrì

Potrei cavarmela cosi..

PERCHÉ L'AI CLINICA È "SPECIALE»?

Alta posta in gioco, eterogeneità, dati imperfetti, contesti mutevoli

- A differenza di altri domini, la clinica è caratterizzata da **eterogeneità biologica e sociale**, dati rumorosi e spesso mancanti, e contesti che cambiano rapidamente: nuovi protocolli, nuovi device, variazioni di prevalenza.
- Questi fattori amplificano il rischio di errore e rendono cruciale la robustezza.
- In letteratura, un tema ricorrente è il **"dataset shift"**: un modello può performare bene in sviluppo ma degradare in deployment.
- Questa asimmetria tra laboratorio e reparto è la prima trappola: se non progettiamo pensando al cambiamento, l'accuratezza diventa effimera e la sicurezza si indebolisce.

L'ambiente clinico richiede approcci specifici che tengano conto della sua natura dinamica e complessa.

Che cos'è il bias (nell'AI per la medicina)?

"Bias" è qualsiasi **distorsione sistematica** che porta un modello a dare risultati **non corretti o ingiusti** in modo prevedibile (non casuale).

Non è un semplice errore !

è un errore **con una direzione**, spesso legato a **come sono stati raccolti i dati**, **come sono state fatte le etichette** o **come viene usato il modello**.

Le principali fonti di bias

La nostra mente è pigra, molto pigra. Per questo motivo è portata a cercare informazioni che confermino quelle già presenti in memoria

- **Bias di campionamento (sampling bias)**
  Il dataset non rappresenta bene la popolazione reale. Esempio: tanti adulti, pochi bambini; molti pazienti di un singolo ospedale; sottorappresentati alcuni fenotipi (es. rinite non-T2).
  → **Effetto:** il modello funziona bene "dove ha visto" e male altrove (altri centri, età, etnie, comorbilità).

- **Bias di misurazione (measurement bias)**
  Le variabili (input) sono misurate in modo diverso tra gruppi/centri (strumenti, protocolli, scale) oppure hanno tanto "rumore".
  → Effetto: il modello impara differenze **strumentali** (es. marca dell'analizzatore) invece che **cliniche**.

- **Bias di etichettatura (label bias / ground truth bias)**
  L'"oro standard" non è davvero oro: diagnosi discordanti tra clinici, criteri eterogenei, outcome proxy.
  → **Effetto:** l'algoritmo apprende etichette imperfette e amplifica incoerenze preesistenti.

- **Bias algoritmico (model/optimization bias)**
  Scelte di modello, loss o regolarizzazione privilegiano la performance **media** penalizzando sottogruppi minoritari.
  → Effetto: AUROC alto, ma **calibrazione scarsa** o **sensibilità bassa** proprio in chi ha più bisogno.

- **Bias di deployment (dataset shift/drift)**
  Cambia il contesto d'uso: nuova prevalenza, diverso mix di pazienti, nuove linee guida o device.
  → Effetto: il modello "invecchia" e sbaglia sistematicamente in nuovi scenari.

A Yale 6 sistemi di AI per allerta precoce in pazienti critici hanno fornito risultati diversi tra loro!!

Che cos'è l'inequità?

L'**inequità** è la **conseguenza clinica** del bias: **differenze ingiuste** di performance e di esito tra sottogruppi (età, sesso, etnia, stato socio-economico, comorbilità, centro).
In pratica: se il modello sbaglia **più spesso** su un gruppo vulnerabile, **aumenta** il rischio di diagnosi tardive, sovra/sottotrattamento, costi inutili e sfiducia.

Bias e inequità
Da problema etico a problema clinico

Problema clinico reale

Il bias algoritmico non è soltanto questione di giustizia sociale: è un **problema clinico**. Campionamenti squilibrati, etichette imprecise e proxy socio-economici impliciti portano a prestazioni peggiori nei sottogruppi già vulnerabili.

Conseguenze duplicate

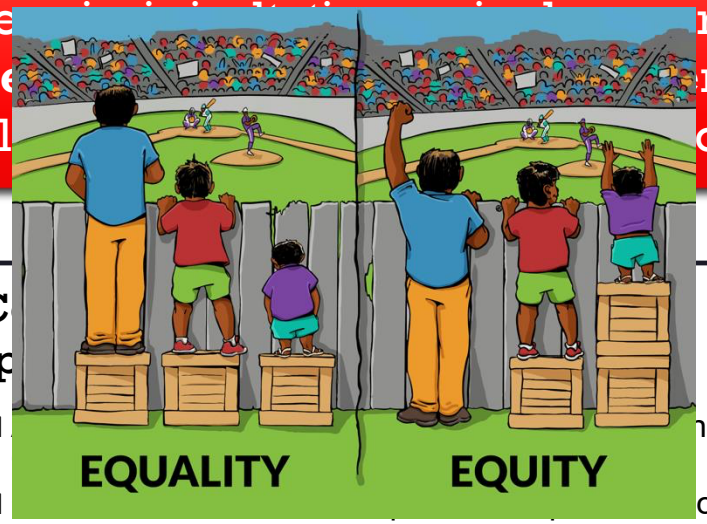
La conseguenza è duplice: **riduzione di efficacia** e **aumento del rischio iatrogeno**. Gli studi recenti raccomandano audit del bias lungo tutta la pipeline: analisi di rappresentatività nei dati, metriche di equità in addestramento e test.

Domanda chiave

La domanda critica non è "quanto bene va il modello?", ma **"per chi e in quali condizioni?"** I piani di mitigazione richiedono arricchimento mirato dei dataset e monitoraggio continuo delle performance nei diversi sottogruppi di pazienti.

Perché è (anche) un problema clinico?

- ❑ Riduce l'**efficacia**: pe rvirebbe di più.
- ❑ Aumenta la **iatrogene** ntrati su alcuni gruppi.
- ❑ Erode la **fiducia** di cl one responsabile



Come si misura l'equita?

Oltre alle metriche globali, si analizzano **metriche per sottogruppo**:

- ❑ **Sensibilità, specificità, PPV/NPV per gruppo** (età, sesso, centro, fenotipo).
- ❑ **Calibrazione per gruppo** (Brier, reliability curves).
- ❑ **Parity/fairness metrics** quando appropriato (es. differenze di tasso positivo).
- ❑ **Net benefit decisionale per gruppo** (analisi decision curve).

C (p

- ❑ nel
- ❑ on coincide col rischio reale in un sottogruppo (es. sovrastima nei bambini piccoli).
- ❑ **Soglie non eque**: una stessa soglia operativa genera troppi falsi negativi in un gruppo.
- ❑ **Proxy sociali occulti**: variabili "innocue" (CAP, orari, percorso clinico) catturano differenze di accesso/risorse, riproducendo disuguaglianze.

Mini esempi pratici

- ❑ **Pazienti al DEA** non necessariamente sono quelli più gravi.
- ❑ **Colture su materiali biologici** eseguite con tecniche diverse(measurement + deployment bias).
- ❑ **LLM per referti**: allucina suggerimenti terapeutici "plausibili" ma non supportati; i medici junior tendono a fidarsi (automation bias) → necessaria policy di double-check.

Qualità metodologica e reporting

Overfitting, coorti piccole, reporting incompleto

Dove inciampiamo più spesso?

a) Overfitting (modello "troppo su misura")

Succede quando il modello impara il rumore del dataset di sviluppo (pattern casuali, artefatti di centro/strumento) e poi **crolla** su dati nuovi.

b) Coorti piccole e campioni "sottili"

Poche osservazioni o pochi **eventi** per quanto predici (es. outcome rari) → stime instabili. Con ML il problema si amplifica, specie con feature ad alta dimensionalità.

c) Reporting incompleto

Studi che non dicono chiaramente **chi** (criteri dei pazienti), **cosa** (predictors esatti e come trattati/mancanti), **come** (feature engineering, tuning), **con quali dati** (prevalenze, split temporali/gerarchici), **con quali metriche** (inclusa **calibrazione**), e **dove** è stato validato (centri/periodi). Con reporting carente non puoi **replicare**, **valutare bias** o **implementare**.

La comunità ha risposto con iniziative di reporting e valutazione strutturate. In pratica, adottare questi standard alza l'asticella, perché obbliga a esplicitare ogni elemento critico del processo di sviluppo e validazione.

TRIPOD-AI

Che cos'è: l'aggiornamento del TRIPOD per modelli predittivi basati su **regressione o ML**. Dice **cosa** riportare perché un lettore possa capire, replicare e giudicare trasferibilità e rischio.

PROBAST-AI

Che cos'è: l'aggiornamento di **PROBAST** per giudicare **rischio di bias** e **applicabilità** di studi/modelli, inclusi quelli **AI/ML**. Lavora su 4 domini (Partecipanti, Predittori, Outcome, Analisi) con domande guidate.

Standard elevati

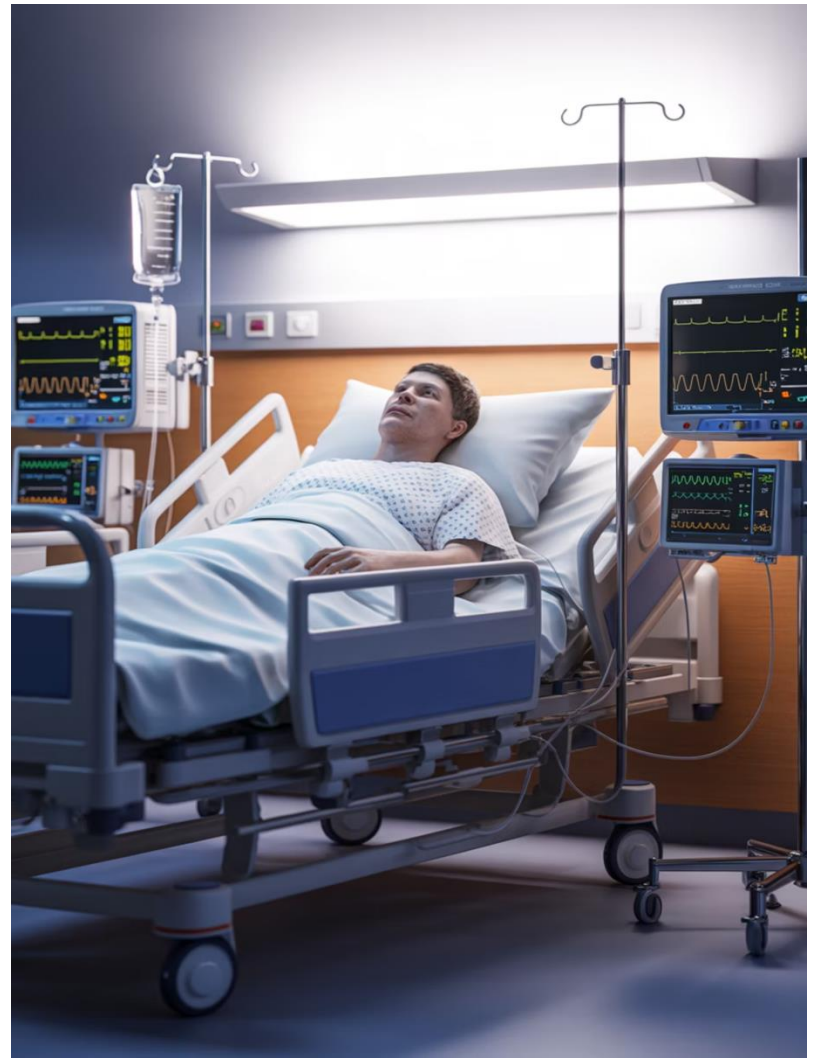
Obbligo di esplicitare popolazione, outcome, strategie di validazione e piani di trasferibilità. aggiungi **DECIDE-AI** per il reporting della **valutazione clinica precoce** (integrazione nel workflow, fattori umani, sicurezza d'uso).

Dalla metrica al letto del paziente

Calibrazione, utilità decisionale, validazione esterna e prospettica

Una metrica elevata, non garantisce valore clinico. In scenari reali contano dimensioni ben più complesse e articolate della semplice accuratezza aggregata.

- **Calibrazione :** La probabilità stimata corrisponde al rischio reale? Essenziale per decisioni basate su soglie.
- **Analisi decisionale:** Net benefit e trade-off: l'impatto sulle scelte cliniche concrete.
- **Validazione esterna:** Performance su siti e periodi diversi, condizione necessaria per la generalizzabilità.
- **Valutazione prospettica:** Contesto reale dove fattori umani e di workflow influenzano l'impatto effettivo.



Dataset shift e drift

Il modello invecchia: monitoraggio continuo e ri-addestramento controllato

Tre tipi di cambiamento

- **Covariate shift:** chi vediamo cambia nel tempo
- **Prior shift:** quanto spesso vediamo che un evento varia
- **Concept drift:** il concetto stesso di outcome si modifica

Questi cambiamenti possono far "invecchiare" un modello anche in pochi mesi, compromettendone l'affidabilità clinica.

MLOps (= *Machine Learning Operations*) sanitario maturo

È l'insieme di processi, pratiche e strumenti che permettono di sviluppare, validare, rilasciare e mantenere in produzione i modelli di ML in modo affidabile, tracciabile e continuo.

- ❑ In produzione serve un monitoraggio continuo: rilevazione di anomalie anche senza etichette rapide, allarmi su spostamenti distribuzionali, e protocolli per ri-addestramento controllato e tracciato.
- ❑ Un MLOps sanitario maturo prevede soglie definite, procedure di roll-back, e documentazione completa delle versioni di dati e modello. L'assenza di questi elementi è una trappola operativa sottovalutata ma critica.

MLOps clinico

Il valore dell'AI non sta nel modello isolato, ma nel **sistema completo** che lo circonda e lo gestisce lungo tutto il ciclo di vita operativo.

Versionamento

Tracciabilità completa di dati, pesi, codice e configurazioni

con rollback controllato

Allarmi su drift

Rilevazione automatica di degradazione e anomalie distribuzionali

Pipeline riproducibili

Automazione e standardizzazione dei processi di sviluppo e deployment

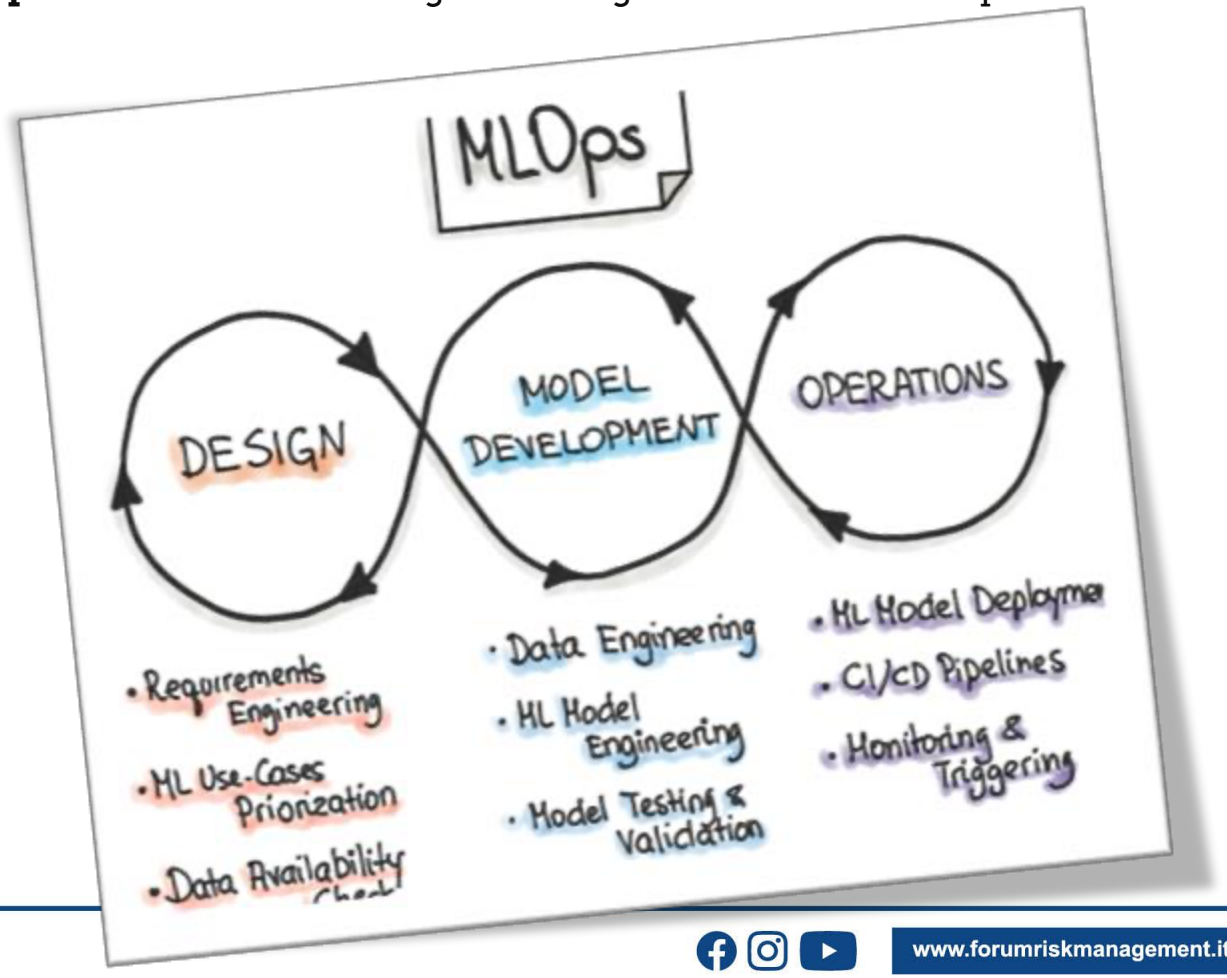
Metriche in tempo reale

Monitoraggio continuo di sicurezza, fairness e performance clinica

Incident response

Protocolli definiti per gestione errori, roll-back e comunicazione

Documentazione completa per accountability e compliance regolatoria

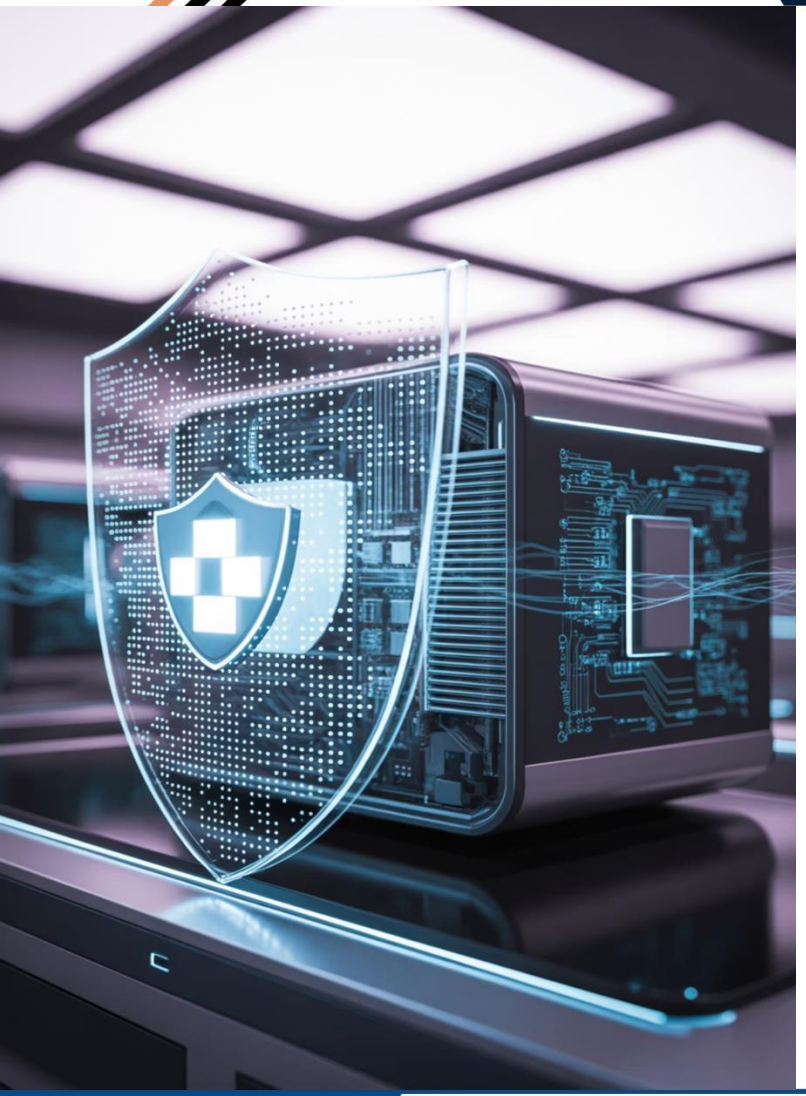


Ma anche un problema economico!!

E' necessaria una "manutenzione/ controllo" sugli algoritmi che possono decadere, con delle spese rilevanti per il Sistema

A Stanford 10 mesi impiegati per verificare 2 modelli di AI

Secondo Coliff (FDA) è molto complessa la verifica della efficienza degli algoritmi



Sicurezza: attacchi avversari e spiegazioni manipolabili

Robustezza ≠ accuratezza media

Perturbazioni avversarie

Sui dati di imaging e segnali, l'AI è vulnerabile a **perturbazioni avversarie** quasi invisibili all'occhio umano che inducono errori sistematici e prevedibili. Questi attacchi possono compromettere diagnosi critiche.

Manipolazione delle spiegazioni

Non sono immuni nemmeno le tecniche di spiegabilità: esistono attacchi che manipolano heatmap e saliency map, minando la fiducia del clinico e la comprensione del processo decisionale.

Le implicazioni sono significative: i test di penetrazione "adversarial" dovrebbero entrare nel collaudo standard dei sistemi AI clinici; e le spiegazioni non vanno trattate come "verità assoluta", ma come strumenti fallibili da validare sperimentalmente. La robustezza richiede valutazione dedicata, separata dall'accuratezza media.

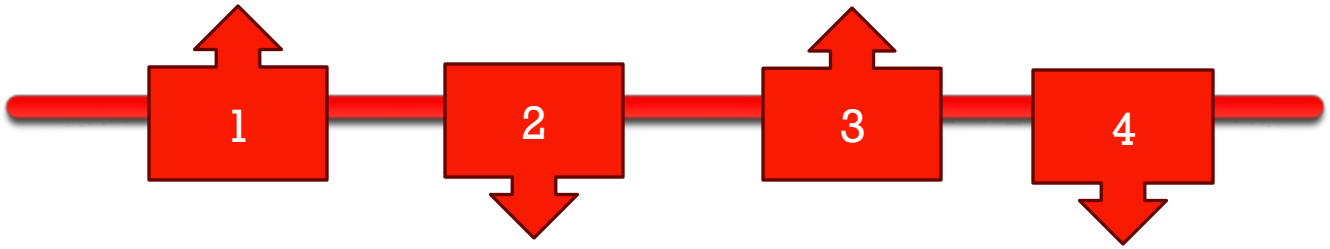
Quadro regolatorio: EU AI Act & MDR/IVDR

High-risk, gestione del ciclo di vita, sorveglianza post-market

In Europa, gran parte dell'AI a uso clinico rientra nella categoria **high-risk** dell'AI Act, con obblighi stringenti su risk management, qualità dei dati e monitoraggio post-market continuo.

AI Act
Requisiti per sistemi high-risk: trasparenza, robustezza, sorveglianza umana

IEC 62304
Standard per il software come dispositivo medico



MDR/IVDR
Regolamenti dispositivi medici e diagnostici in vitro

Integrazione
Mappatura congiunta dei requisiti sovrapposti

Questo si sovrappone ai requisiti dei dispositivi medici (MDR/IVDR) e agli standard software (per es., IEC 62304). In pratica, occorre una **mappatura congiunta**: evidenza clinica proporzionata all'uso previsto, tracciabilità completa dei dataset, trasparenza del modello e piani di sorveglianza strutturati. La non allineata integrazione regolatoria è una trappola frequente nei progetti che passano dal laboratorio al mercato.

Checklist "anti-trappole”:Otto dimensioni per un'AI clinica affidabile

1) **Dati**
Audit di rappresentatività, analisi di bias nei campioni, piani di arricchimento mirato per sottogruppi vulnerabili

2) **Metodo**
Reporting strutturato TRIPOD-AI, valutazione del rischio di bias con PROBAST-AI, gestione rigorosa dei dati mancanti

3) **Valutazione**
Calibrazione verificata, validazioni esterne multi-sito e multi-temporali, valutazione prospettica guidata da DECIDE-AI

4) **Sicurezza**
Test di robustezza avversaria, validazione sperimentale delle spiegazioni, protezione contro manipolazioni

5) **LLM**
Grounding/RAG con fonti verificate, policy di double-check clinico obbligatorio, logging completo delle fonti utilizzate

6) **Privacy**
Federated learning quando appropriato, misure di privacy tecnica (differential privacy, secure aggregation), policy per dati sintetici

7) **Regolatorio**
Mappatura integrata AI Act + MDR/IVDR, evidenza clinica proporzionata, piani di sorveglianza post-market strutturati

8) **MLOps**
Monitoraggio continuo multi-dimensionale, procedure di roll-back testate, incident response documentato, audit trail completo

Conclusioni

AI utile = AI sicura + valida + governata

Le trappole dell'AI in medicina sono oggi ben documentate: bias e inequità, scarsa trasferibilità tra contesti, vulnerabilità di sicurezza multiple, rischi specifici dei Large Language Models, sfide di privacy nella condivisione dei dati, e requisiti regolatori complessi e non banali da soddisfare.

La **buona notizia** è che esistono strumenti e pratiche mature per mitigarle in modo sistematico ed efficace:

- ❑ **Reporting trasparente** con standard TRIPOD-AI e valutazione PROBAST-AI
- ❑ **Validazioni esterne e prospettiche** guidate da framework come DECIDE-AI
- ❑ **MLOps sanitario** con monitoraggio continuo, incident response e audit trail
- ❑ **Governance dei dati** con federated learning e misure di privacy tecnica
- ❑ **Allineamento regolatorio** integrato tra AI Act, MDR/IVDR e standard software

Affidabile nel tempo

Robusta al dataset shift e ai cambiamenti del contesto clinico

Equa tra i pazienti

Performance verificate nei sottogruppi vulnerabili e minoritari

Integrata nei processi

Utilizzabile dai clinici con supporto al workflow e alla decisione

L'AI diventa veramente clinica quando soddisfa questi tre criteri simultaneamente. Questo è il ponte tra promessa tecnologica e valore concreto per la salute dei pazienti. L'innovazione responsabile richiede rigore metodologico, trasparenza, e governance continua lungo tutto il ciclo di vita del sistema.

GRAZIE PER L'ATTENZIONE

Prof. Francesco Macrì