

10-11
OTTOBRE
2019
MILANO
PALAZZO
LOMBARDIA



HTA per l' Intelligenza Artificiale: da dove partiamo?

Ing. Matteo Ritrovato, PhD

Responsabile Servizio Health Technology Assessment

IRCCS Ospedale Pediatrico Bambino Gesù

Cos'è l'Intelligenza Artificiale?

*L'intelligenza può essere definita come la capacità di **apprendere o comprendere o di affrontare nuove situazioni o di applicare conoscenze o abilità per manipolare il proprio ambiente.***



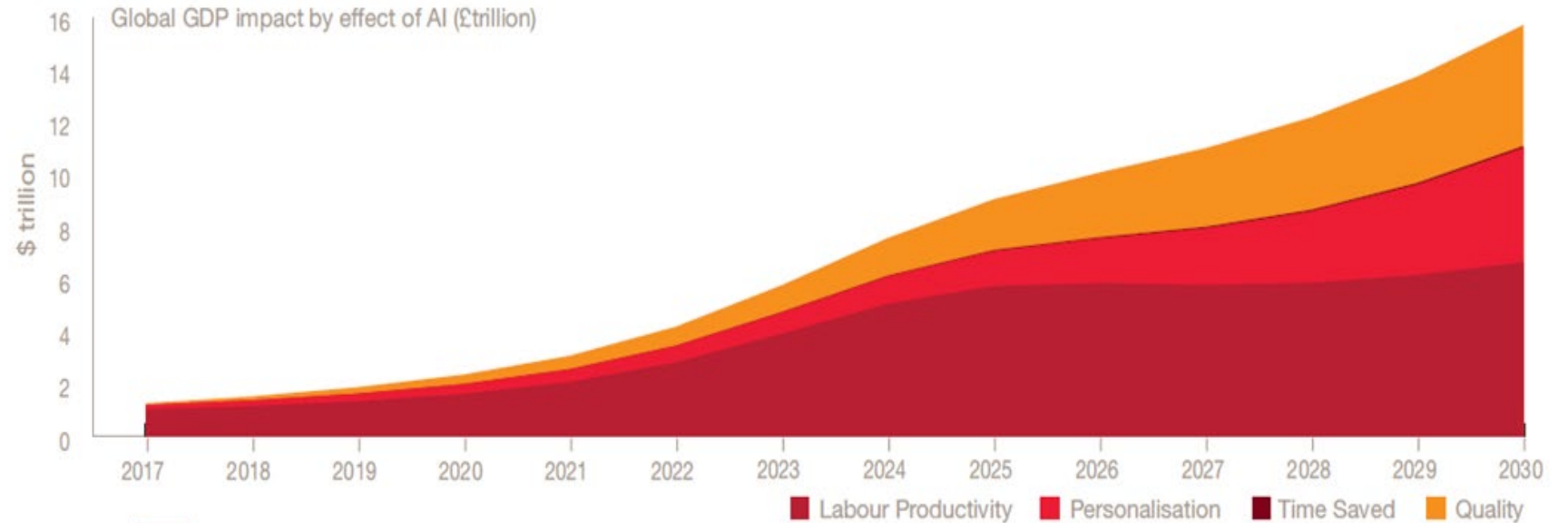
*L'intelligenza artificiale (AI) è la simulazione dei processi di intelligenza umana da parte delle macchine, in particolare dei sistemi informatici. Questi processi includono **l'apprendimento** (acquisizione di informazioni e regole per l'utilizzo delle informazioni), il **ragionamento** (usando le regole per raggiungere conclusioni approssimative o definite) e **l'auto-correzione**.*

Prospettive sull'AI

Figure 1: Where will the value gains come from with AI?

- ❑ L'AI è chiaramente il futuro in molti settori produttivi
- ❑ È impensabile escludere la medicina dai settori impattati dall'AI

(fonte PwC)

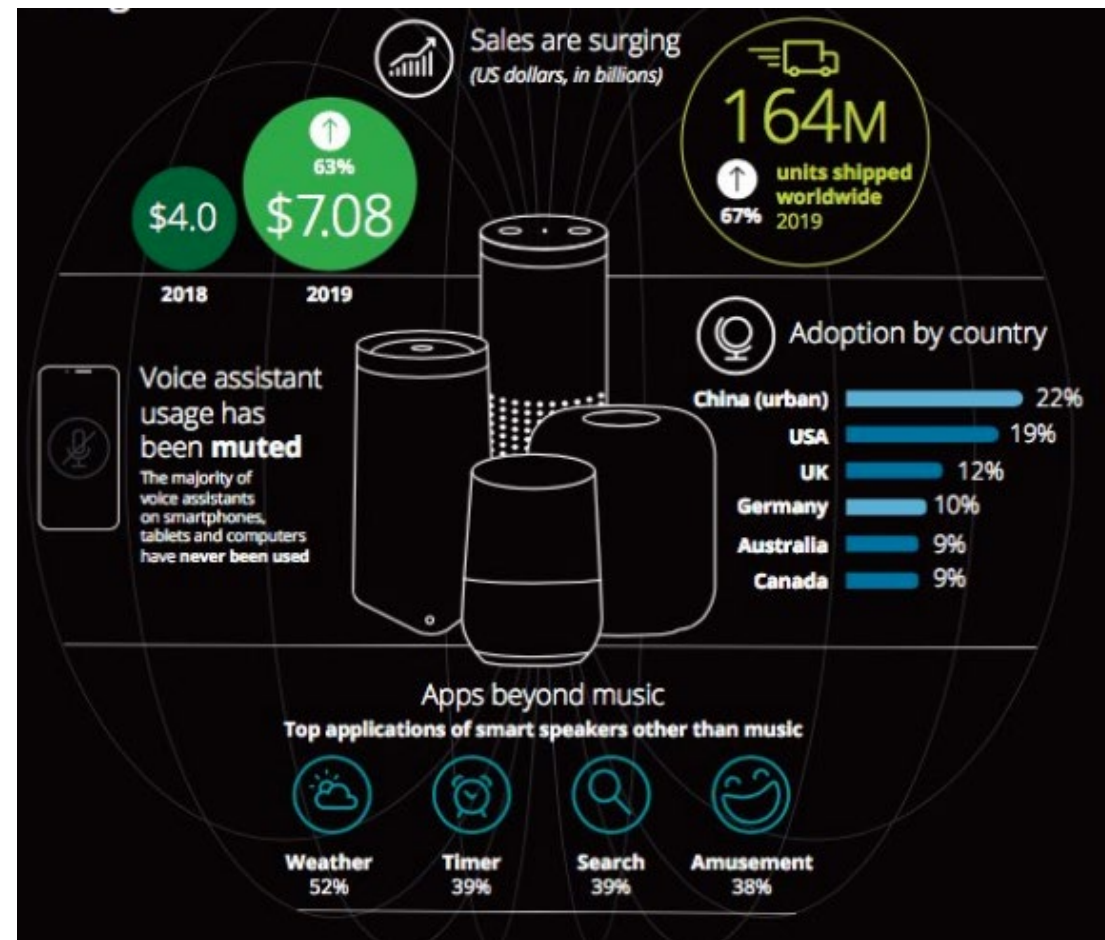


Source: PwC analysis

Al attorno a noi

- **Smart speaker:** mercato stimato al valore **7 miliardi** di dollari nel 2019
- **Veicoli guida autonoma:** potrebbero arrivare in strada **entro il 2020**, Il mercato potrebbe arrivare a valere **54,23 miliardi di dollari** entro la fine del 2019 (Allied Market Research)
- **Profilazione utenti** (per motivi commerciali, politici, etc.): *il “valore di mercato” dell’utente medio è intorno ai 289.19 dollari per anno*

La spesa mondiale in soluzioni Ai dovrebbe aggirarsi attorno ai **46 miliardi di dollari entro i prossimi due anni**, passando per i 19 miliardi di fine 2018 e i quasi 30 miliardi di fine 2019, per poi raggiungere i **57,6 miliardi di dollari nel 2021**.



AI in Medicina

- Primo articolo pubblicato sull' AI nel campo della biomedicina nel 1961:

Heuristic glimpses into the evolution of radiation therapy; particularly in reference

to gynecologic cancers. LEUCUTIA T. Am J Roentgenol Radium Ther Nucl Med. 1961 Jan;85:3-20.

- Da allora relativa scarsità di studi pubblicati su riviste mediche e una conseguente mancanza di interesse tra i clinici sull'applicazione dell'AI in medicina:
 - l'IA e temi correlati sono stati **11.225** / 4.419.743 o lo **0,25%** del totale delle pubblicazioni relative alla medicina).
- Nel 2016 **solo 1.274** circa i rapporti sulle applicazioni dell'intelligenza artificiale in medicina, **su un totale di quasi 408.000** articoli trovati su PubMed, corrispondenti allo **0,31%**.
- Usando una terminologia più specifica come «**machine learning**», il numero di pubblicazioni diminuisce in modo significativo a circa **813** (o lo **0,20%** delle pubblicazioni totali); (nel 2006 solo un totale di **20 risultati** con la parola chiave «machine learning» **su un totale di 146.697** articoli in medicina (o **0,01%**).

AI in Medicina

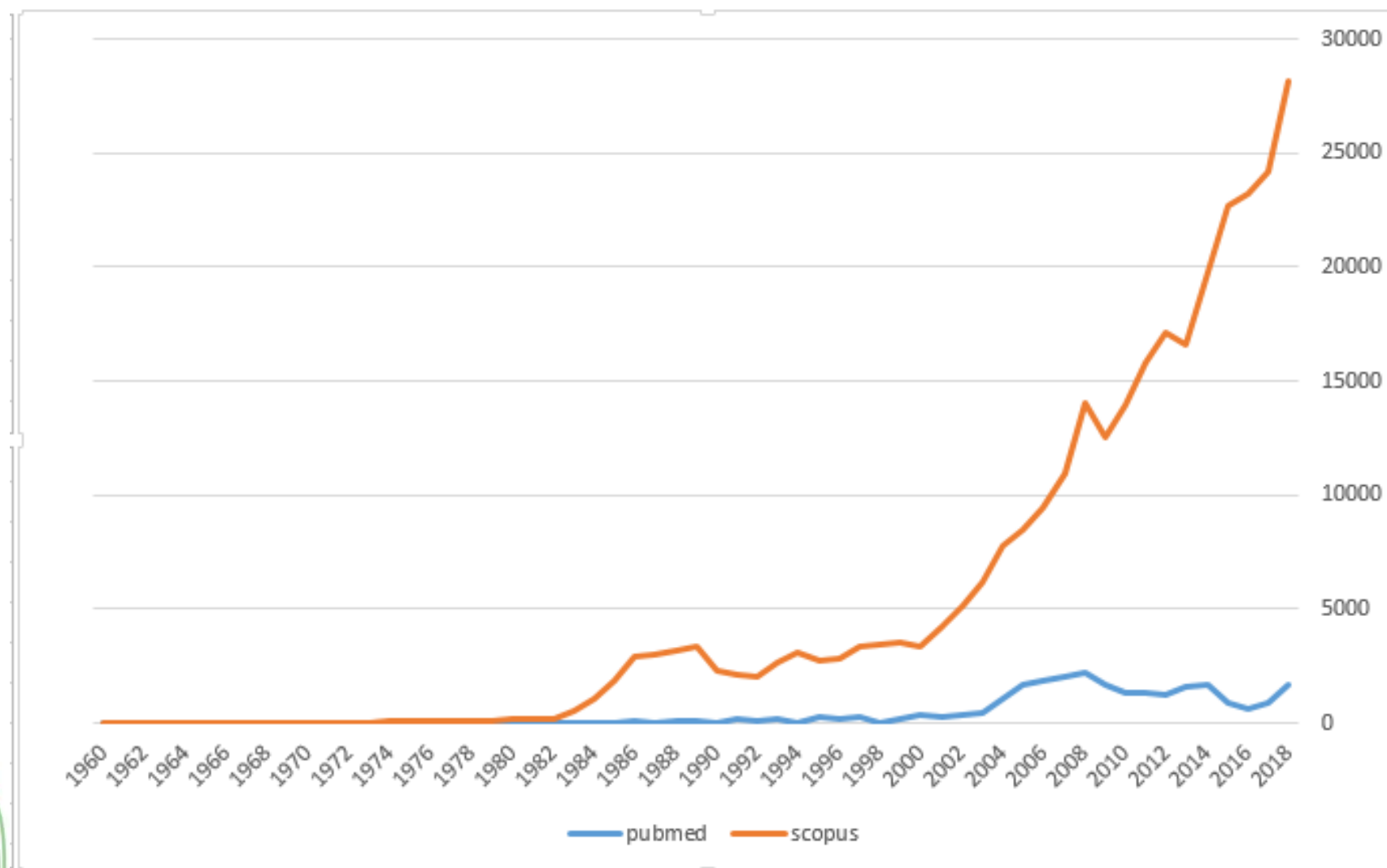
The state of AI hype has far exceeded the state of AI science, especially when it pertains to validation and readiness for implementation in patient care. (Nature medicine, 2019)

Investimenti privati (VC) nel 2017 pari 1.3Miliardi di US\$ (cumulativo dal 2014 pari a 2.3 miliardi di US\$) *(fonte KPMG, 2018)*

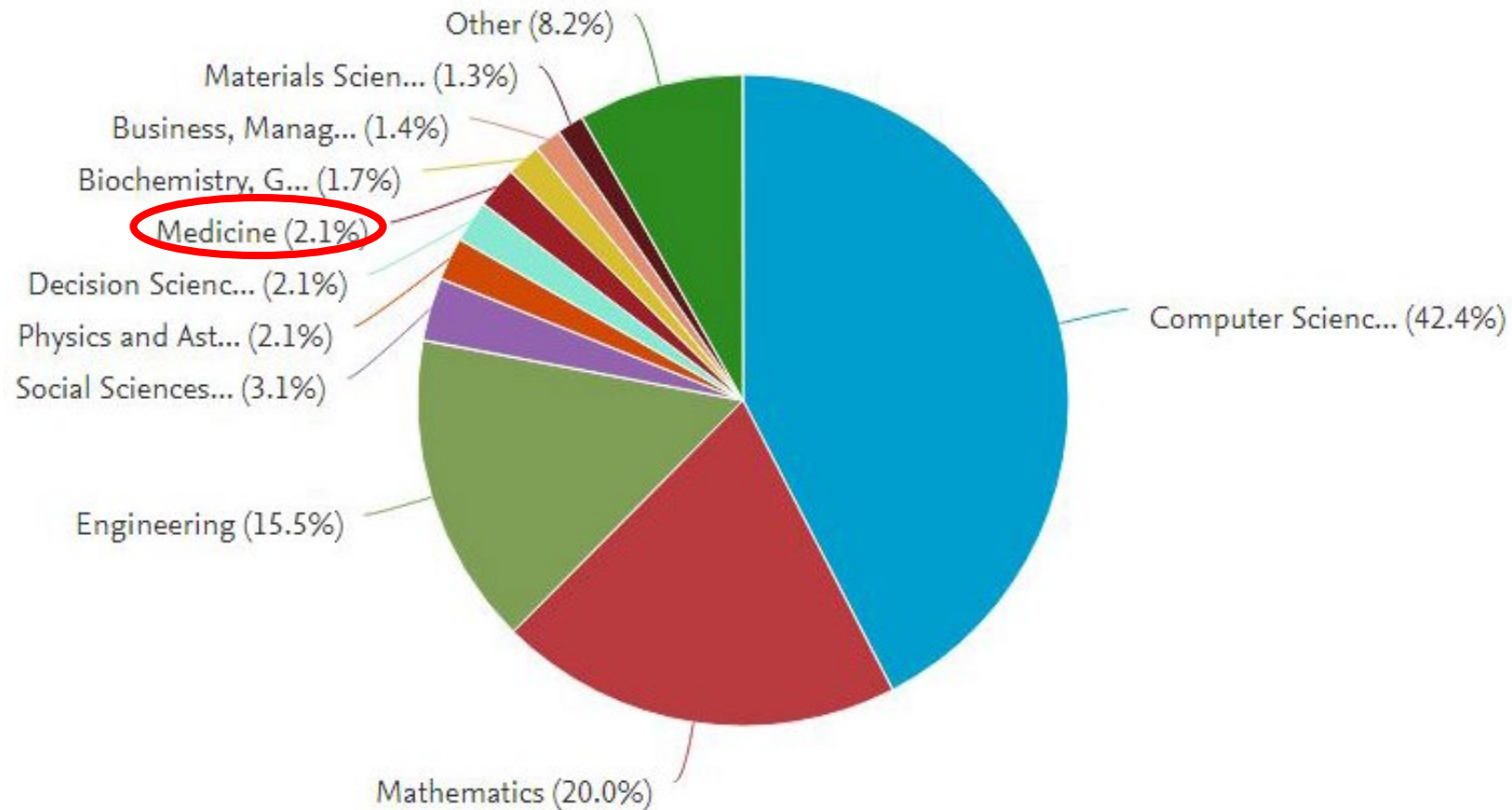
Investimenti complessivi, pubblici e private, raggiungeranno i 6.6 miliardi di US\$ nel 2021. *(Forbes insight, 2019)*

Le principali applicazioni di AI in sanità produrranno risparmi ai sistemi sanitari pari 150 miliardi di US\$ all'anno entro il 2026 *(Accenture, 2019)*.

Produzione Scientifica



Produzione scientifica: gli ambiti



La Letteratura

Perché AI in medicina?

- L'uso maggiore dell'AI in campi come l'oncologia (ad esempio per l'oncogenetica) o l'epidemiologia (previsioni sull'incidenza/prevalenza di patologie) può essere spiegato dalla necessità di gestire ed estrapolare un outcome da un'enorme quantità di dati.
- L'uso in radiologia, invece, ha una triplice predisposizione intrinseca:
 1. Analogo, enorme, quantitativo di dati
 2. introduzione del digitale molto anteriore rispetto ad altri ambiti clinici (es cartella clinica elettronica)
 3. sfruttamento di tecnologie già mature (primi algoritmi di *pattern recognition* risalgono agli anni '70).

Subspecialty	Total Published Reports (since 1950)	Published Reports 2006	Published Reports 2016
High Use (>1,000)			
Oncology	9,998	411	660
Pathology	6,354	224	459
Radiology	4,541	255	264
Epidemiology	2,595	99	194
Neurosurgery	1,011	74	41
Critical Care Medicine	1,006	41	66
Medium Use (250-1,000)			
Psychiatry	859	25	120
Nuclear Medicine	809	43	77
Internal Medicine	745	28	67
Neurology	715	20	110
Obstetrics and Gynecology	669	8	51
Population Health	650	27	80
Cardiology	447	18	52
Pediatrics	364	8	58
Anesthesiology	351	6	27
Low Use (<250)			
Ophthalmology	214	8	24
Emergency Medicine	203	4	28
Dermatology	140	4	21
Orthopedic Surgery	133	6	14
Endocrinology	120	2	16
Neonatal Medicine	107	3	16

Imaging & Diagnostica

- Vi sono molte promesse nell'utilizzo di tecniche come il deep learning per l'interpretazione automatica e/o aumentata dell'immagine biomedica in radiologia, patologia, dermatologia, oftalmologia e cardiologia nel prossimo futuro.

JAMA | **Original Investigation** | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

EXPOSURE Deep learning-trained algorithm.



← Editorial

+ Supplemental content

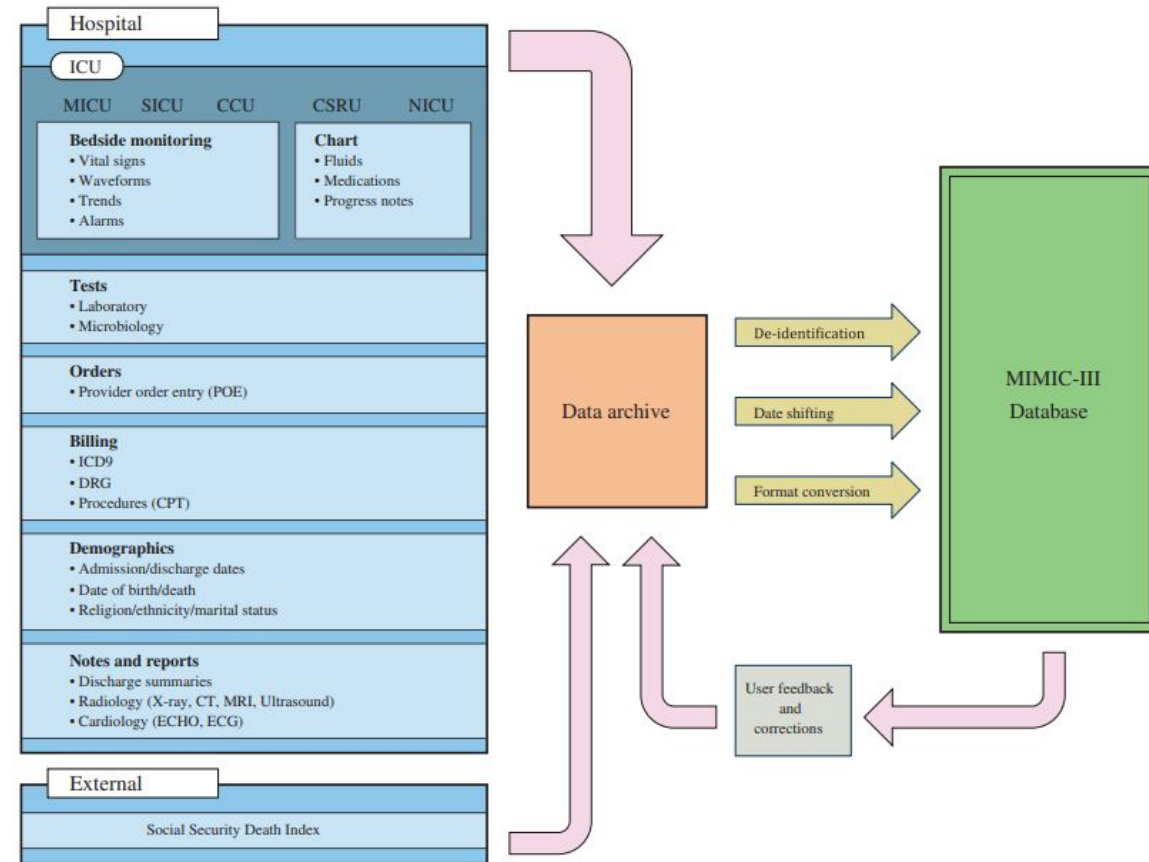
Medicina di precisione

- La medicina di precisione con la sua complessità e l'enormità di dati da analizzare è particolarmente adatta per metodologie AI come il deep learning per identificare e valutare pazienti con patologie simili.



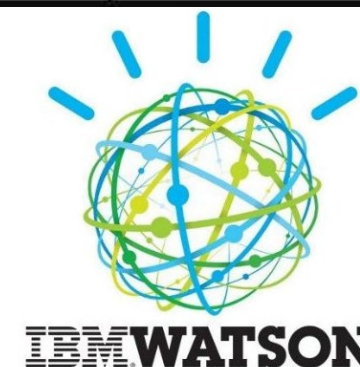
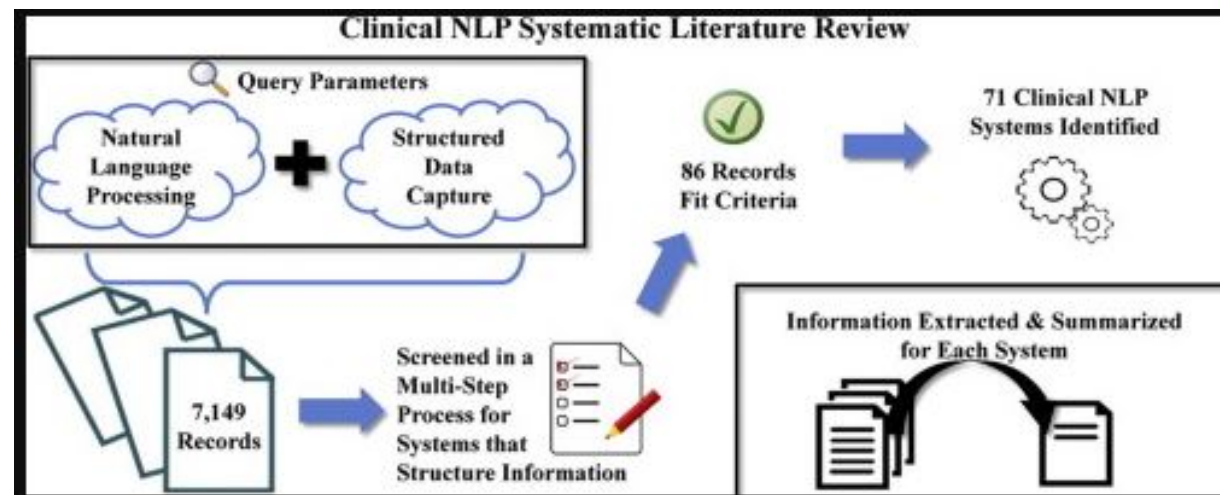
Decision Support and Hospital Monitoring

- Una recente review della storia del «Clinical Decision Support» (CDS) afferma il notevole miglioramento in questo settore grazie all'avvento di ausili cognitivi a supporto di diagnosi, terapia, coordinamento dell'assistenza, sorveglianza e prevenzione, mantenimento della salute o del benessere
- Chiavi di ricerca pubmed e risultato numerico ricerca:
 - «Clinical Decision Support» AND «error rate» -> 27 risultati
 - «Clinical Decision Support» AND «adverse events» -> 7 risultati

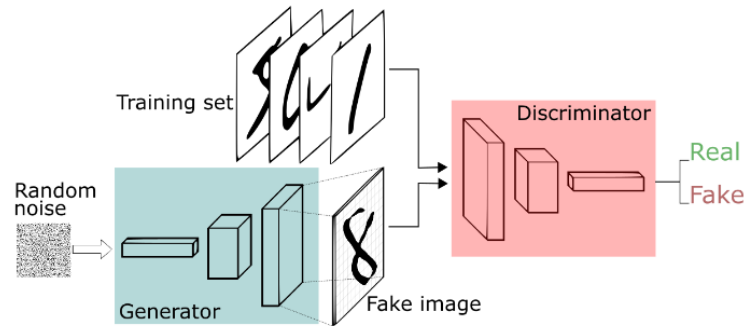


AI per il continuo aggiornamento del clinico

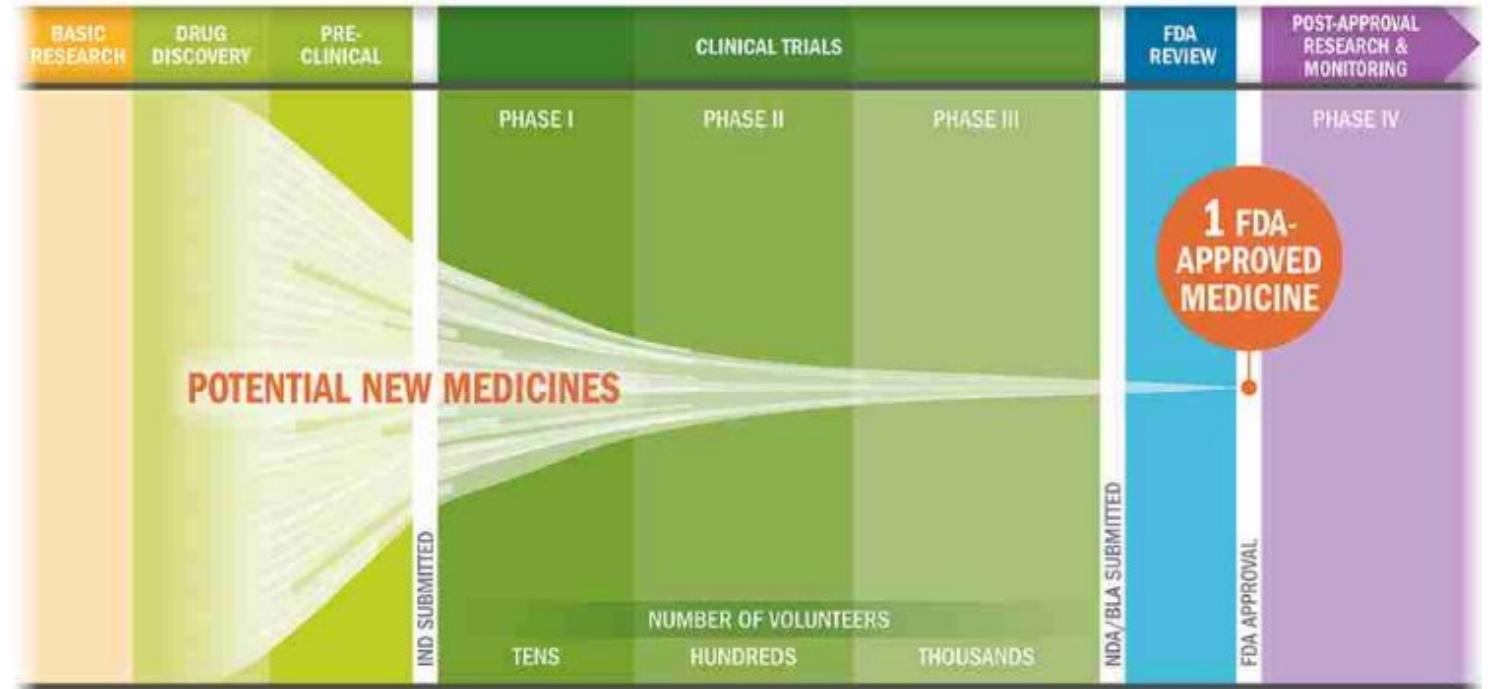
Il numero di articoli scientifici pubblicati è raddoppiato ogni nove anni dalla seconda guerra mondiale, e ora ammonta a oltre **3 milioni di articoli sottoposti a revisione ogni anno** da oltre 34.000 riviste accademiche.



Nuovi Farmaci



Generative Adversarial Network



Costo stimato per R&D di ogni singola molecola arrivata sul mercato pari a 2,6 miliardi di US\$

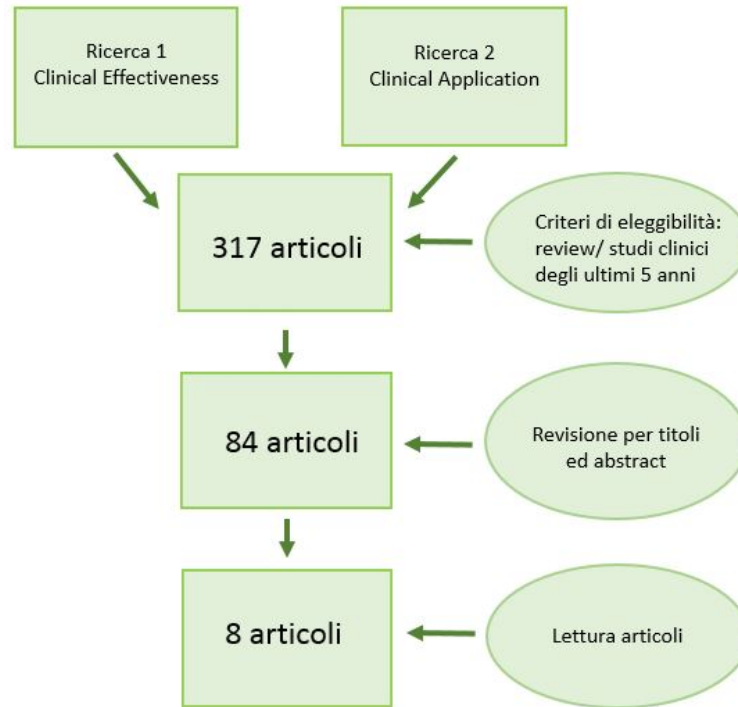
(J.A. DiMasi, et al., *Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics.* 47, 2016)

Revisione di Letteratura

Efficacia Clinica

Chiavi di ricerca:

- Ricerca 1: «artificial intelligence» AND «clinical effectiveness»
- Ricerca 2: «artificial intelligence» AND «clinical outcome»



Ambito di applicazione degli 8 articoli:

- Management and depression prediction
- Speech recognition per persone con disabilità nel parlare
- Perdita di peso
- Stress management
- Trattamento della sindrome dell'arto fantasma
- Smettere di fumare
- Nutrizione personalizzata sulla base della risposta glicemica

Nonostante il considerevole **aumento di letteratura** nell'uso dell'AI nella ricerca clinica, è evidente come **pochissimi di questi studi vengano realmente tradotti nella pratica clinica.**

Revisione di Letteratura

- Su oltre 20.000 articoli revisionati meno dell'1% hanno un design sufficientemente robusto.
- Solo 25 studi hanno validato i dati esternamente (usando immagini di una diversa popolazione)
- Solo 14 studi comparavano la performance dell'AI con il medico specialista, usando lo stesso campione

A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdass, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

Summary

Background Deep learning offers considerable promise for medical diagnostics. We aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.

Methods In this systematic review and meta-analysis, we searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019. Studies comparing the diagnostic performance of deep learning models and health-care professionals based on medical imaging, for any disease, were included. We excluded studies that used medical waveform data graphics material or investigated the accuracy of image segmentation rather than disease classification. We extracted binary diagnostic accuracy data and constructed contingency tables to derive the outcomes of interest: sensitivity and specificity. Studies undertaking an out-of-sample external validation were included in a meta-analysis, using a unified hierarchical model. This study is registered with PROSPERO, CRD42018091176.

Findings Our search identified 31587 studies, of which 82 (describing 147 patient cohorts) were included. 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy, with sensitivity ranging from 9.7% to 100.0% (mean 79.1%, SD 0.2) and specificity ranging from 38.9% to 100.0% (mean 88.3%, SD 0.1). An out-of-sample external validation was done in 25 studies, of which 14 made the comparison between deep learning models and health-care professionals in the same sample. Comparison of the performance between health-care professionals in these 14 studies, when restricting the analysis to the contingency table for each study reporting the highest accuracy, found a pooled sensitivity of 87.0% (95% CI 83.0–90.2) for deep learning models and 86.4% (79.9–91.0) for health-care professionals, and a pooled specificity of 92.5% (95% CI 85.1–96.4) for deep learning models and 90.5% (80.6–95.7) for health-care professionals.

Interpretation Our review found the diagnostic performance of deep learning models to be equivalent to that of health-care professionals. However, a major finding of the review is that few studies presented externally validated results or compared the performance of deep learning models and health-care professionals using the same sample. Additionally, poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy. New reporting standards that address specific challenges of deep learning could improve future studies, enabling greater confidence in the results of future evaluations of this promising technology.

Revisione di letteratura

Chiara assenza di studi sperimentali sugli outcome:

- ✓ Clinici
- ✓ Organizzativi
- ✓ Economici

Emerge anche l'assenza di evidenze relative alle **modalità di indagine e misura** di tali outcome

Cosa è chiaro sinora: Impatto sul lavoro

- diminuzione carico lavoro (utilizzo assistivo)
 - In radiologia la performance dell'AI varia da patologia a patologia (AUC score che varia da 0.63 per diagnosi di polmonite a 0.87 per collasso dei polmoni o ingrossamento del cuore)
 - In dermatologia si arriva ad AUC pari al 0.96
 - aiuto nei compiti ripetitivi (utilizzo sostitutivo)
- Potrà avere un impatto anche sull'organizzazione dei sistemi?



Cosa è chiaro sinora: Big Data ins't enough

- Qualità del dato: non conta solo «quanti» ma anche e soprattutto «quali» e «come»
- Anche per l'AI vale il detto Garbage In Garbage Out
- In assenza di adeguati ed interconnessi sistemi informativi, l'AI non potrà mantenere le aspettative



Cosa è chiaro sinora: High Performance Medicine



- la disponibilità di sistemi software che «ingeriranno» ed «elaboreranno semanticamente» grandi quantità di dati (anche in forma testuale) in modo rapido, accurato ed economico, in grado di condensarne e restituire il significato, quindi in grado di vedere e svolgere attività altrimenti non (umanamente) possibili (cfr **high performance medicine**, *Nature Medicine* 2019).
- ambito di utilizzo su cui convergono le maggiori aspettative



sihta

Cosa è chiaro sinora: Aspetti Etici, Legali e Sociali

- ✓ Protezione dei Dati (GDPR)
- ✓ Responsabilità legale (nuovi regolamenti e linee guida FDA e UE)
- ✓ Rapporto medico/paziente (etica e accettabilità sociale)
- ✓ Assicurazione sanitaria (rischio della profilazione utenti)



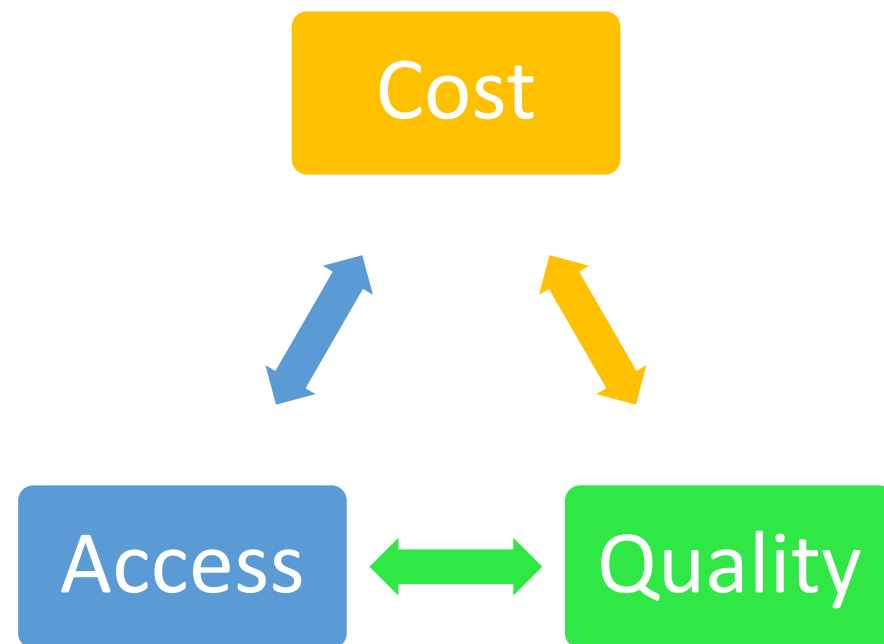
Cosa si può fare in termini di HTA

- Definizione dei modelli e metodi di valutazione
- Definizione degli outcome principali (chiari, specifici, ben circoscritti) da sperimentare
- Definizione delle metriche con cui misurare gli outcome

Domain	LV1	LV2	
Technology	Algorithms	stability	
		sensitivity/specificity	
		accuracy	
		calculation speed	
	Data	Data collection	
		Data bias	
Clinical Effectiveness	Radiology	Radiomics	number of features
			features selection method specificity
			lesion segmentation accuracy
		images interpretation	classify patients
			integrating information
			disease detection
	Genomics/Genetics	Pattern recognized in DNA	
		Genetic causes of a disease	
		gene expression	
	Pathology	Tissue classification	
		rare target identification	
	Oncology	survival rate (5-10-30), overall survival rate/ prognosis	
		treatment response prediction	
		early diagnosis	
	Chronic disease	diabetes treatment: efficacia nel controllo emoglobina glicosilata	
		Depression treatment: % guarigione	
		preventing readmission	
		prescription efficacy	
Safety	Preventing medical error		
Costs	indirect costs	installation	
		staff training	
		labour cost savings	
		software update	
		maintenance	
	direct costs	purchase	
Legal, Social and Ethical aspects	Regulations		
	Data protection		
	Clinical Legal Responsibility		
	Effects of AI on patients		
Organizational aspects	Process	Level of Process automation vs health operators workload	
		increased production	
		waiting list reduction	

Conclusioni

- Enormi aspettative attorno all'AI
- Soluzione al dilemma del «triangolo di ferro» (W. Kissick, 1994)
- Verificare che il miglioramento degli outcome economici/organizzativi (**minori costi, maggiore produttività, maggiore accessibilità**)) sia accompagnato da un analogo miglioramento dell'outcome clinica, per via diretta (**migliori terapie**) o indiretta (es. **maggiore accessibilità**→**cure più tempestive**→**cure migliori**)



Conclusioni

- Identificare modelli valutativi e misure di outcome chiare ed univoche da sperimentare
 - Outcome può essere di tipo clinico, oppure di tipo economico o organizzativo
- Riemerge chiaramente il problema della misura dell'outcome clinico in condizioni non adeguate per l'RCT (cfr medicina di precisione (singolo paziente/caso unico), outcome incrementale/marginale su milioni di pazienti/cittadini,) e quindi di come produrre misure di efficacia alternative (in silico trials, modellistica, analisi del principio di funzionamento, etc.)

Conclusioni

La high performance medicine sembra invece la frontiera più vicina, perché in qualche modo più comprensibile:

1- Da un lato perché "semplicemente" una reale «***data based medicine***» (non più basata, solamente, sulle capacità mnemoniche, logiche, ed intellettive dei professionisti sanitari)

2- dall'altro perché l'AIA (***Artificial Intelligence Augmentation***) ha già dimostrato (in altri settori) le proprie potenzialità e semplicità attuativa

In assenza di dimostrazioni sull'efficacia clinica, organizzativa e/o economica, si limiterebbe ad un nuovo strumento per la formazione e l'educazione continua dei professionisti o per la ricerca scientifica

Conclusioni

*When you create a human + AI system,
the hard part ins't the "AI".
It isn't even the "human".
It's the "+"*

(Case N. Journal of Design and Science, MIT, 2018)

Grazie per l'attenzione.

Ing. Matteo Ritrovato, PhD
Responsabile Servizio di Health Technology Assessment
IRCCS Ospedale Pediatrico Bambino Gesù
matteo.ritrovato@opbg.net

