Recenti Prog Med 2017; 108: 397-401

Potenziali conseguenze inattese dell'uso di sistemi di intelligenza artificiale oracolari in medicina

FEDERICO CABITZA[1,2*], CAMILLA ALDERIGHI[3*], RAFFAELE RASOINI[3,4*], GIAN FRANCO GENSINI[3*]

[1]IRCCS Istituto Ortopedico Galeazzi, Milano; [2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università di Milano-Bicocca, Milano; [3]CESMAV - Centro Studi Medicina Avanzata, Firenze; [4]IFCA Istituto Fiorentino di Cura e Assistenza, Firenze.
*Florence EBM/Renaissance Group.

Pervenuto il 16 agosto 2017. Accettato dopo revisione il 31 agosto 2017.

Riassunto. I sistemi di supporto decisionale basati sul *machine learning* (ML) in medicina stanno raccogliendo un crescente interesse grazie a recenti pubblicazioni che ne hanno evidenziato l'elevata accuratezza diagnostica in specifici contesti clinici. Tuttavia, agli ipotetici vantaggi derivanti dall'applicazione dei sistemi di intelligenza artificiale in campo medico, vanno criticamente affiancati alcuni potenziali inconvenienti. Alla luce dell'attuale mancanza di studi sugli effetti collaterali dell'applicazione di questi nuovi supporti decisionali nella pratica medica, in questo articolo sintetizziamo le principali conseguenze inattese che potrebbero derivare dalla loro applicazione estesa, soprattutto in relazione alla peculiare caratteristica dei sistemi di tipo "oracolare", ovvero che si basano su modelli predittivi di ML in cui l'elevata accuratezza spesso è inversamente proporzionale alla trasparenza del percorso che conduce alle predizioni. Le principali conseguenze inattese correlate con l'impiego dei sistemi di ML variano dall'incertezza intrinseca dei dati impiegati per "addestrare" e alimentare questi sistemi, all'inadeguata esplicabilità delle loro risposte, al rischio di una eccessiva tendenza ad affidarsi a questi sistemi da parte dei loro utenti, con una loro conseguente dequalificazione e desensibilizzazione nei confronti del contesto clinico. Sebbene alcune di queste criticità possano essere difficili da valutare, data la scarsità a oggi di reali applicazioni di tali sistemi nella pratica medica, vogliamo richiamare l'attenzione su di esse al fine di dare un contributo di indirizzo per ricerche future e in supporto di politiche di approvazione dei sistemi di intelligenza artificiale che siano maggiormente informate e consapevoli, e che sappiano vedere oltre le valutazioni, spesso eccessivamente enfatiche, che riguardano questo tipo di tecnologia informatica a supporto della decisione medica.

"Handle with care": about the potential unintended consequences of oracular artificial intelligence systems in medicine.

Summary. Decisional support systems based on *machine learning* (ML) in medicine are gaining a growing interest as some recent articles have highlighted the high diagnostic accuracy exhibited by these systems in specific medical contexts. However, it is implausible that any potential advantage can be obtained without some potential drawbacks. In light of the current gaps in medical research about the side effects of the application of these new AI systems in medical practice, in this article we summarize the main unexpected consequences that may result from the widespread application of "oracular" systems, that is highly accurate systems that cannot give reasonable explanations of their advice as those endowed with predictive models developed with ML techniques usually are. These consequences range from the intrinsic uncertainty in the data that are used to train and feed these systems, to the inadequate explainability of their output; through the risk of overreliance, deskilling and context desensitization of their end-users. Although some of these issues may be currently hard to evaluate due to the still scarce adoption of these decisional systems in medical practice, we advocate the study of these potential consequences also for a more informed policy of approval beyond hype and disenchantment.

Introduzione

Recentemente, l'applicazione del *machine learning* (ML) in medicina, ossia di quell'approccio alla realizzazione di sistemi di intelligenza artificiale a supporto delle decisioni in cui buona parte della ottimizzazione dei modelli decisionali è affidata all'"apprendimento automatico"[1], ha fatto grandi passi avanti e sta suscitando un interesse crescente: il numero delle pubblicazioni indicizzate su Medline relative all'impiego di tali sistemi, per esempio, è aumentato di circa dieci volte nell'ultima decade rispetto al periodo precedente (figura 1).

In particolare, l'applicazione di una tecnica di ML detta di "apprendimento profondo" (*deep learning*) ha portato alcuni sistemi informatici a esibire un'accuratezza diagnostica comparabile a quella di medici esperti in diversi campi, quali la diagnosi di retinopatia diabetica[2] e quella di tumori dermatologici[3].

Tuttavia, se i lavori finora pubblicati hanno indagato soprattutto l'accuratezza di questi sistemi, ancora mancano in letteratura studi sull'efficacia del loro impiego in rapporto a obiettivi clinici importanti, come la riduzione della mortalità o il miglioramento della qualità di vita dei pazienti.

Numerosi sono i vantaggi che si prospettano derivare dall'applicazione del ML alla medicina: dall'au-

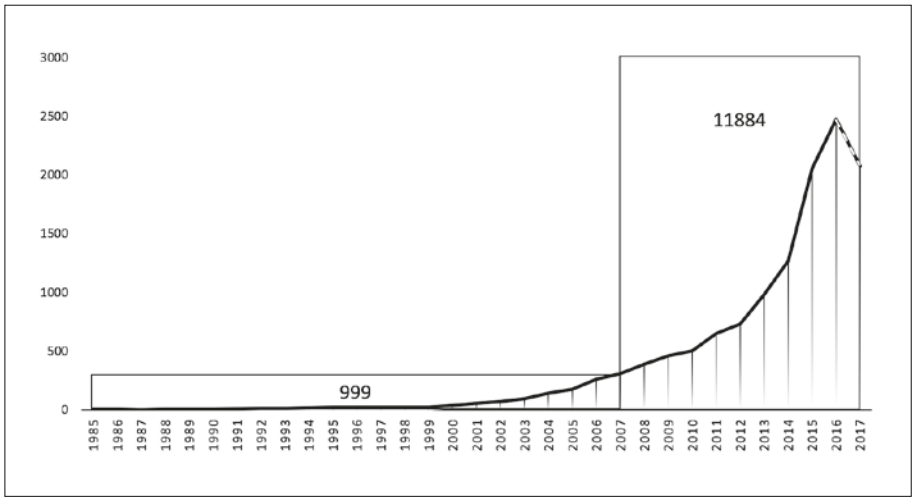

Figura 1. Numero di articoli in Medline su applicazioni machine learning.

mento della produttività del sistema, alla maggiore accuratezza diagnostica, fino alla possibilità di favorire l'accesso agli esami diagnostici anche in luoghi e a persone che non possono beneficiarne a causa di barriere geografiche, politiche ed economiche[4,5].

Tuttavia, gli effetti derivanti da qualsiasi tipo di innovazione devono essere considerati con un approccio conservativo, se non costruttivamente critico. Nell'ambito delle numerose pubblicazioni relative all'introduzione dei ML-DSS (*decision support systems*) in medicina, in prevalenza favorevoli, ve ne sono alcune che invece mirano a far luce sui limiti attuali e sulle possibili conseguenze inattese di tali sistemi[6]. Il nostro gruppo di ricerca si colloca tra questi ultimi con la recente pubblicazione su *JAMA* di un viewpoint sull'argomento[7]. Il nostro e contributi simili non si pongono l'obiettivo di fornire risposte certe, considerato lo scarso impiego, attualmente, di ML-DSS in contesti reali, ma piuttosto quello di sensibilizzare i decisori politici, i direttori di struttura, i dirigenti medici e tutti gli utilizzatori finali rispetto a un esercizio di inquadramento del problema che sia ben fondato e che si articoli con domande che è opportuno considerare per capire come sia meglio procedere in un terreno ancora essenzialmente sconosciuto. Per esempio, è opportuno chiedersi se tutti gli attori coinvolti accetteranno l'introduzione di questi strumenti; come li utilizzeranno nella pratica quotidiana; fino a che punto alcuni ruoli professionali saranno ridisegnati e come si modificheranno le relazioni tra medici e il rapporto medico-paziente; quali effetti comporterà l'impiego del ML in termini economici, di responsabilità medico-legale e di formazione didattica; su quali obiettivi clinici sarà misurata l'efficacia di questi strumenti decisionali, come precondizione necessaria per la loro approvazione e adozione su larga scala; se sarà accettabile integrare nella propria pratica medica raccomandazioni di sistemi quali quelli di deep learning, generate per mezzo di percorsi non trasparenti tanto agli occhi dei medici quanto a quelli degli esperti di ML, come ammesso candidamente anche da questi ultimi[8].

Nell'ottica di contribuire a indirizzare la ricerca verso alcune possibili risposte, e nel contempo per superare il dibattito, probabilmente sterile, sull'accuratezza crescente dei sistemi di ML (dibattito sterile se a confrontarsi è sempre il sistema automatico preso isolatamente, cioè non integrato in un team di specialisti bensì messo in competizione con gli specialisti stessi), ci focalizziamo su alcune possibili conseguenze inattese dell'impiego di tali sistemi decisionali, ipotizzandoli comunque sempre "al meglio" della loro funzione, ossia nella condizione ideale in cui la loro accuratezza di classificazione sia equivalente (se non superiore) rispetto a quella umana.

Criticità correlate all'impiego di sistemi oracolari in medicina

Una delle maggiori criticità correlate con l'integrazione dei ML-DSS nella pratica medica è costituita dal rischio che nel tempo i medici possano sviluppare un ingiustificato ed eccessivo affidamento nelle capacità dell'automazione (*over-reliance*)[9]. Questa fiducia sarebbe alimentata dal convincimento che ogni nuova tecnologia sia intrinsecamente migliore di qualsiasi altra già in uso e che, una volta che una operazione sia stata automatizzata, il supporto tecnologico debba essere considerato al pari o migliore di ogni essere umano coinvolto nello stesso incarico[10].

In un contesto a elevata complessità e incertezza decisionale come quello medico, è ragionevole ipotizzare che potrebbe verificarsi un ricorso notevole all'automazione, qualora essa dimostrasse un'elevata affidabilità predittiva, anche come conseguenza di pratiche di medicina difensiva ancora da immaginare.

Purtroppo, come conseguenza dell'eccessivo affidamento, vi è il rischio di sviluppare una vera e propria dipendenza (*overdependence*) da questi sistemi che, nel lungo periodo, potrebbe condurre alla dequalificazione (*deskilling*), ovvero alla riduzione del livello di competenza richiesto per svolgere una fun-

zione, quando tutte o alcune delle componenti dei compiti corrispondenti siano state automatizzate. Il fenomeno del deskilling diverrebbe più evidente nel momento in cui la tecnologia fallisse o cessasse di funzionare, anche solo temporaneamente, e sarebbe non meno insidioso quando utilizzatori resi meno competenti e quindi meno abili nell'ottimizzare i modelli predittivi dei ML-DSS dovessero rendere l'evoluzione dei loro strumenti e quindi il loro miglioramento continuo più lento o più difficile[11].

La letteratura offre già alcuni esempi di questo fenomeno: in un'analisi condotta da parte di un gruppo di ricercatori della City University of London sulla lettura di 180 mammogrammi da parte di 50 professionisti, è stata documentata una riduzione della sensibilità diagnostica del 14,5% per il rilievo di cancro mammario nei medici più esperti, quando a questi venivano presentate immagini di difficile lettura corredate con l'interpretazione da parte del computer, mentre solo un aumento dell'1,6% della sensibilità diagnostica è stato rilevato grazie al supporto del computer nel sottogruppo di medici meno esperti quando a questi venivano presentati casi di più semplice interpretazione[12]. Questi risultati sottolineano non solo come l'eccessivo affidamento nei sistemi di ML da parte degli operatori possa influenzarne la performance, ma anche che molta ricerca è necessaria per individuare le dinamiche di questo fenomeno, soprattutto in rapporto alla diversa esperienza dei medici coinvolti e alla diversa difficoltà dei casi loro presentati.

Un altro potenziale aspetto critico da considerare nell'ambito dell'introduzione dei ML-DSS in medicina riguarda la possibile progressiva sottovalutazione del contesto, spesso di difficile rappresentazione ed espressione esplicita, rispetto a ciò che invece è facilmente codificabile ed esprimibile a parole o per mezzo di numeri, ovvero rispetto ai dati, che sono necessari perché qualsiasi ML-DSS possa funzionare e risultare utile.

I ML-DSS, alimentati da un numero finito di dati discreti, e incapaci di incorporare elementi poco o per niente "datificabili", quali per esempio aspetti culturali, sociali o psicologici di un paziente, oppure aspetti organizzativi di un contesto ospedaliero, potrebbero indurre i medici a trascurare la funzione di questi elementi che invece sappiamo essere necessari per un'accurata ed efficiente gestione del percorso di cura e perni insostituibili dell'irripetibile interazione medico-paziente[13].

Un esempio in cui l'abilità predittiva di un ML-DSS è risultata tecnicamente valida, ma potenzialmente fuorviante, è stato illustrato da Caruana et al.[8] nell'ambito di una casistica di 14.199 pazienti con polmonite, su cui sono stati valutati differenti algoritmi di ML allo scopo di predire il rischio di mortalità e indirizzare così la gestione dei pazienti in ambito intra- o extra-ospedaliero. Gli algoritmi analizzati in tale studio evidenziarono che i pazienti con storia di asma erano a minore rischio di morte rispetto ai pazienti non asmatici. Questa indicazione sorprese i ricercatori, i quali esclusero che l'asma potesse rappresentare un fattore protettivo in tali pazienti e attribuirono l'"errore" a un cosiddetto "intervento medico confondente"[14]: negli ospedali coinvolti nello studio, infatti, i pazienti con polmonite e storia di asma erano solitamente ricoverati in terapia intensiva e presentavano una minore mortalità, probabilmente in virtù di un maggiore controllo clinico e di una gestione aggressiva della patologia. Dunque, per la mancanza di una variabile dicotomica nei dati relativa all'ammissione del paziente in terapia intensiva, quest'ultimo elemento non poteva essere considerato dagli algoritmi valutati: ciò illustra il caso paradossale in cui algoritmi formalmente perfetti possano sbagliare proprio in virtù della loro perfetta aderenza ai dati, i quali sono invece spesso incompleti, caratterizzati da elementi di incertezza, o persino (in una certa ma importante misura) inaccurati[15].

La criticità di includere fattori di difficile rappresentazione nei processi decisionali basati su algoritmi di ML, e l'impossibilità di escludere da essi ogni fattore confondente, potrebbe condurre a similari errori contestuali, che un'eccessiva fiducia e dipendenza nella tecnologia potrebbero rendere più frequenti o difficili da rilevare ed evitare.

Un ulteriore elemento da analizzare è l'incertezza, che caratterizza intrinsecamente ogni fenomeno in medicina[16]. La ricerca attuale sui ML-DSS, che è come noto condotta da una prospettiva principalmente teorica e ingegneristica, pare ignorare questo aspetto così pervasivo della pratica medica e trascurarne l'impatto sulla validità e attendibilità dei dati con cui ogni ML-DSS è programmato nelle fasi di "addestramento", validato nelle fasi di test e verifica, e quindi alimentato durante l'uso operativo[17].

Il punto fondamentale è che i dati necessari all'algoritmo per essere addestrato ed elaborare le sue predizioni sono solitamente forniti dall'essere umano: per esempio, per istruire un algoritmo nel riconoscimento di una certa patologia, come il melanoma[3], a partire da immagini diagnostiche, il ML-DSS riceve in ingresso centinaia (o migliaia) di fotografie, nell'esempio suddetto fotografie di lesioni cutanee e, più in generale, casi già corredati da diagnosi "corrette", cioè precedentemente classificate da parte di specialisti.

Questo meccanismo presenta due principali punti deboli, relativamente alla validità dei dati e alla loro attendibilità.

Il primo aspetto riguarda la discrepanza tra qualità dei dati usati per l'addestramento del sistema e qualità dei dati nelle cartelle cliniche informatizzate attraverso cui si ritiene che gli ML-DSS elaboreranno le proprie predizioni nel prossimo futuro: mentre la qualità dei dati di addestramento è spesso elevata, grazie al necessario sforzo di ripulitura e pre-elaborazione che sarebbe insostenibile garantire anche nella pratica clinica routinaria, la qualità dei cosiddetti "real-world data" soddisfa raramente i presupposti di "dato ideale" che dovrebbe alimentare un algoritmo di ML[15,17].

Pertanto, di fronte a ML-DSS in grado di elaborare milioni di dati con elevati livelli di accuratezza, una

Non è possibile, a oggi, pensare ai ML-DSS come ipotetici sostituti del medico e nemmeno come dispositivi in grado di migliorarne le prestazioni.

domanda da porsi riguarda la qualità dei dati stessi, poiché la scarsa qualità dell'input di qualsiasi algoritmo esita in un output inaffidabile, in accordo con la nota espressione "garbage-in, garbage-out".

Il secondo elemento di potenziale criticità riguarda l'attendibilità dei dati che è altresì legata alla intrinseca e in buona parte ineludibile incertezza dei fenomeni osservati in medicina. Come è noto, nell'ambito degli studi sull'accuratezza di nuovi test diagnostici, si denota come "gold standard" un test impiegato come riferimento rispetto al quale diviene misurabile l'accuratezza di un secondo test diagnostico che si intende valutare, così da stabilire una possibile "verità" di riferimento. Tuttavia, per numerose diagnosi, esistono non uno, ma più gold standard di apparente pari utilità[17]. Nel recente lavoro condotto da ricercatori di Google sulla diagnosi di retinopatia diabetica[2], per esempio, il gold standard impiegato per addestrare l'algoritmo nella rilevazione di immagini retiniche patologiche era la decisione maggioritaria di un gruppo di oftalmologi. Lo studio ha dimostrato il raggiungimento di elevati livelli di accuratezza diagnostica da parte del ML-DSS utilizzato; ciò nonostante, alcuni autori hanno sostenuto che, se fosse stata impiegata come gold standard la tomografia ottica a coerenza di fase, come fatto in altri studi relativi alla retinopatia diabetica, ciò avrebbe potuto modificare i livelli di accuratezza ottenuti[18].

Inoltre, anche se esistesse un unico standard di riferimento accettato da gran parte della comunità medica, è stato evidenziato che il livello di accordo tra osservatori sullo stesso esame diagnostico di riferimento può risultare tutt'altro che ottimale[19,20].

Il rischio di considerare attendibile (cioè indipendente dal valutatore e dall'osservatore) l'associazione tra un insieme di dati e una specifica diagnosi non può essere sottovalutato da alcun medico che sia seriamente interessato al supporto dei ML-DSS nella sua pratica, a causa della rigidità con cui questi sistemi producono le loro predizioni, ossia sulla base di associazioni, ritenute "vere per definizione", ma in realtà codificate, più o meno arbitrariamente, sulla base di indicazioni, anche discordanti, di uno o più osservatori. In queso caso il rischio peggiore è la possibile instaurazione di un circolo vizioso in cui i pattern a cui sono sensibili gli ML-DSS sono evidenziati e suggeriti agli specialisti medici e questi divengono meno sensibili a identificarne altri oppure gli stessi ma autonomamente. Questo fenomeno, a cui abbiamo dato il nome di "sclerosi epistemica", può da un lato mettere a rischio la capacità dei sistemi di ML di evolversi e migliorare progressivamente in un contesto soggetto a progressivi mutamenti e nuove acquisizioni come quello medico e, dall'altro, rischia di cristallizzare nel modello predittivo alla base del ML-DSS correlazioni basate su dati "sporchi", oppure interpretazioni caratterizzate da elevata variabilità inter-osservatore.

Conclusioni

Per condividere alcune raccomandazioni tanto sul piano metodologico quanto su quello pragmatico relativamente all'impiego del ML in medicina, sottolineiamo l'importanza, nell'ambito degli studi di validazione, di porre attenzione sia alla qualità dei dati con cui i modelli predittivi sono definiti e validati, sia alla qualità dei dati su cui tali modelli operano nell'ambito della loro applicazione reale. In particolare, auspichiamo che si dedichi maggiore ricerca sulla valutazione degli ambiti di produzione e di consumo dei dati clinici e sulla identificazione e riduzione dei relativi errori contestuali.

Dovrebbe poi divenire ben radicata, tra gli informatici e tra gli esperti di ML, l'esigenza di non trascurare l'incertezza insita nell'interpretazione di ogni fenomeno in medicina e anzi prioritario considerare tale incertezza, e valorizzarla, nell'ambito degli algoritmi di definizione dei modelli predittivi e nella interpretazione dei loro risultati[17].

Andrebbe inoltre evidenziato e studiato il pericolo di sovra-affidamento e di eccessiva dipendenza da sistemi di accuratezza che definiamo "oracolare", cioè molto elevata ma non associata a spiegazioni esplicite e significative per gli operatori coinvolti[8], con gli effetti secondari di deskilling e desensibilizzazione al contesto clinico.

Infine, vogliamo auspicare una maggiore diffusione di studi clinici sull'applicazione del ML che confrontino team di specialisti "umani" che si avvalgono delle indicazioni fornite loro da ML-DSS rispetto a singoli individui o team che non sono supportati da alcuna tecnologia di quel tipo, e che non si limitino a indagare misure di accuratezza, che rimangono endpoint surrogati, ma indirizzino obiettivi di importanza primaria, come la mortalità, la qualità di vita e il rapporto tra i costi e il raggiungimento di questi obiettivi.

Sulla base di quanto detto, non è possibile, a oggi, pensare ai ML-DSS come ipotetici sostituti del medico e nemmeno come dispositivi in grado di migliorarne le prestazioni. Sebbene quest'ultimo aspetto sia verosimile e anche auspicabile, la letteratura è stata finora povera di riferimenti in merito agli aspetti su cui abbiamo posto la nostra attenzione in questo articolo; aspetti che, sebbene difficili da valutare al momento, rappresentano le conseguenze inattese più temibili di una adozione indiscriminata degli "oracoli artificiali" in medicina.

Conflitti di interesse: gli autori dichiarano l'assenza di conflitti di interesse.

Bibliografia

1. Cabitza F, Banfi G. Machine Learning in laboratory medicine: waiting for the flood? CCLM 2017; in press. DOI 10.1515/cclm-2017-0287.
2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402-10.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115-8.
4. Obermeyer Z, Emanuel EJ. Predicting the future: Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016; 375: 1216-9.
5. Jha S, Topol EJ. Adapting to artificial intelligence radiologists and pathologists as information specialists. JAMA 2016; 316: 2353-54.
6. Chen JH, Asch SM. Machine Learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017; 376: 2507-9.
7. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in medicine. JAMA 2017; 318: 517-8.
8. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2015; 1721-30.
9. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2011; 19: 121-7.
10. Greenhalgh T. Five biases of new technologies. Br J Gen Pract 2013; 63: 425.
11. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf 2016; 3: 119-31.
12. Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. Med Decis Making 2013; 33: 98-107.
13. Weiner SJ, Schwartz A. Contextual errors in medical decision making: overlooked and understudied. Acad Med 2016; 91: 657-62.
14. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annual Symposium Proceedings 2013; 2013: 1109.
15. Ahmad FS, Chan C, Rosenman MB, et al. Validity of cardiovascular data from electronic sources:the multi-ethnic study of atherosclerosis and HealthLNK. Circulation 2017; pii: CIRCULATIONAHA.117.027436.
16. Simpkin AL, Schwartzstein RM. Tolerating uncertainty: the next medical revolution? NEJM 2016; 375: 1713-5.
17. Cabitza F, Ciucci D, Rasoini R. A giant with feet of clay: on the validity of the data that feed machine learning in medicine. 2017; ArXiv preprint arXiv:1706.06838. URL: https://arxiv.org/abs/1706.068
18. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA 2016; 316: 2366-7.
19. Braun R, Gutkowicz-Krusin D, Rabinovitz H, et al. Agreement of dermatopathologists in the evaluation of clinically difficult melanocytic lesions: how golden is the 'gold standard'? Dermatology 2012; 224: 51-8.
20. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K; Thai Screening for Diabetic Retinopathy Study Group. l. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. Ophthalmology 2006; 113: 826-32.

Indirizzo per la corrispondenza:
Dott. Ing. Federico Cabitza
IRCCS Istituto Ortopedico Galeazzi
Via Riccardo Galeazzi 4
20161 Milano
E-mail: cabitza@disco.unimib.it