Apports de l'intelligence artificielle dans la prise de décision en médecine

The potential of artificial intelligence in medical decision making

GRÉGOIRE J.-M.[1,2], GILON C.[1], CARLIER S.[2] et BERSINI H.[1]

[1]IRIDIA (Artificial Intelligence Research Laboratory), Université libre de Bruxelles (ULB)

[2]Département de Cardiologie, Université de Mons (UMons)

RÉSUMÉ

L'intelligence artificielle est appelée à jouer un rôle de plus en plus important en médecine. Il semble utile de faire un point concernant son utilisation dans la prise de décision médicale. Celle-ci nécessite plusieurs étapes comportant anamnèse, examen clinique, établissement d'un diagnostic différentiel probabiliste, réalisation de tests, confirmation du diagnostic et mise en route du traitement.

Nous comparons les performances de l'homme et de la machine à chacune de ces étapes. Le cerveau humain procède par raisonnement, inférences logiques ou par automatismes généralement acquis par l'expérience. Ces mêmes deux modes décisionnels coexistent en ce qui concerne la machine. Les algorithmes de type raisonnement, basés sur les connaissances, utilisent un mode de fonctionnement similaire aux humains mais leur mise en œuvre est plus complexe et ils exigent une collaboration avec les experts et/ou l'exploitation ingénieuse de la littérature médicale. Ils utilisent des systèmes à base de règles, des représentations structurées de la connaissance (type ontologie), des arbres décisionnels et/ou des modèles graphiques. Pour la seconde famille d'algorithmes, non basés sur les connaissances, des résultats impressionnants ont été obtenus par des réseaux de neurones (notamment dans leur version dite « profonde »), essentiellement dans l'interprétation des tests diagnostiques comportant un traitement d'images ou de signal comme des séries chronologiques.

Certains problèmes restent à résoudre : la reproductibilité des études, la généralisation des modèles et l'accessibilité à des bases de données fiables. L'utilisation des techniques d'intelligence artificielle pour l'optimalisation des prises en charge des patients reste encore un défi.

Rev Med Brux 2022 ; 43 : 265-273
Doi : 10.30637/2022.22-013

ABSTRACT

Artificial intelligence is set to play an increasingly important role in medicine. It seems useful to review its use in medical decision-making. It requires several steps including history taking, clinical examination, establishment of a probabilistic differential diagnosis, doing tests, diagnosis confirmation and treatment initiation.

We compare the performance of humans and machines at each of these stages. The human brain proceeds by reasoning, logical inferences or by automatisms generally acquired through experience. These same two decision-making modes co-exist in the case of the machine. Reasoning-type algorithms use a similar mode of operation to humans, but their implementation is more complex, and they require collaboration with experts and/or ingenious exploitation of the medical literature. They use rule-based systems, structured knowledge representations (ontology type), decision trees and/or graphical models. For the second family of algorithms (non-knowledge based), impressive results have been obtained by neural networks (especially in their so-called «deep» version), mainly in the interpretation of diagnostic tests involving image or signal processing such as time series.

Some problems remain to be solved: the reproducibility of studies, the generalization of models and the accessibility of reliable databases. The use of artificial intelligence techniques to optimize patient care is still a challenge.

Rev Med Brux 2022 ; 43 : 265-273
Doi : 10.30637/2022.22-013

Key words : artificial intelligence, decision making, decision trees, neural networks

INTRODUCTION

L'intelligence artificielle (IA) est définie par le *High Level Expert Group* de la Commission européenne comme un domaine scientifique qui s'intéresse à la compréhension informatique de ce que l'on appelle communément le comportement intelligent et à la création d'agents intelligents qui présentent un tel comportement[1]. Une recherche dans Google comportant les mots *artificial intelligence* et *medicine* rapporte169 millions de sites à consulter et PubMed propose 38.456 articles, montrant une croissance exponentielle. La mise en œuvre de l'intelligence artificielle dans la pratique clinique est un domaine de développement prometteur, qui évolue rapidement avec les autres domaines modernes que sont la médecine de précision, la génomique et la téléconsultation[2-4]. Il semble donc utile de faire un point concernant son utilisation dans le raisonnement médical. La prise de décision médicale nécessite plusieurs étapes incluant anamnèse, examen clinique, établissement d'un diagnostic différentiel probabiliste, réalisation de tests paracliniques, confirmation du diagnostic et mise en route du traitement.

BUT

Comparer les méthodes pratiquées par le médecin avec les techniques d'intelligence artificielle pour l'établissement d'un diagnostic médical suivi de la mise en œuvre d'un traitement.

METHODES

A partir d'une sélection représentative de publications sur l'IA, nous réalisons une comparaison entre les performances de l'homme et de la machine à chaque étape de la procédure utilisée par le médecin pour aboutir au traitement le plus efficace pour le patient : mode de raisonnement, anamnèse, examen clinique, tests, diagnostic final et choix du traitement.

RESULTATS

Modes décisionnels

Pour l'homme, comme pour la machine, plusieurs modes décisionnels coexistent. Le cerveau humain utilise essentiellement deux types de procédure pour décider : les automatismes acquis par expérience et le raisonnement logique. Kahneman évoque d'une manière imagée l'existence de deux systèmes de pensée : le *système 1*, automatique et inconscient, qui permet des prises de décisions rapides mais s'avère sujet à de nombreux biais et le *système 2*, rationnel, logique, plus difficile à mettre en œuvre, se déclenchant pour pallier les ratés du premier, mais moins sujet aux erreurs[5,6].

On utilise le terme *hypothético-déductif de confirmation* pour résumer l'ensemble du processus de prise de décision rationnel du médecin. Ce processus est long et exige des efforts considérables. Pour cette raison le cerveau humain procède souvent par automatismes, des raccourcis de pensée que l'on emprunte lorsque les exigences d'une tâche cognitive, à force de répétitions, ont imprimé à ce point les synapses neuronales qu'elles ne nécessitent plus de recourir aux processus séquentiels de nature consciente. Klein a particulièrement étudié ce phénomène et a mis en évidence l'élément dominant : la reconnaissance inconsciente d'un schéma déjà vécu, basé sur l'expérience. Il a modélisé ce processus dans un algorithme, le *Recognition-Prime Decision model*[7]. Le problème principal de ces automatismes réside dans les biais cognitifs qui en résultent, tel le bien connu biais de confirmation.

Deux approches coexistent en ce qui concerne la machine. L'approche cognitive ou logique s'inspire des processus cognitifs typiquement humains (absents chez les animaux) lorsqu'ils se trouvent confrontés à des situations inattendues. Newell et Simon ont créé le *General Problem Solver,* un programme qui résolvait des problèmes en comparant les étapes de son raisonnement à ceux des humains confrontés aux mêmes problèmes[8]. Les systèmes cognitifs tentent d'imiter les aspects de la pensée humaine, tout en ajoutant la capacité de traiter de grandes quantités d'informations et de grands ensembles de données. Ils procèdent par des processus d'observation (intégration des données sous quelque format que ce soit), d'interprétation (dictionnaires permettant une compréhension spécifique du langage technique), d'évaluation (établissement de l'ensemble de toutes les connexions possibles facilitant la génération d'hypothèses) et de décision (facilitée par la restructuration des données et la mise en évidence des preuves).

Cette approche, logique ou symbolique, s'appuie sur l'idée que nous raisonnons en appliquant des inférences logiques de type « SI… ALORS… ». Elle utilise des systèmes d'arbres décisionnels, des graphes de résolution et même la logique floue[9]. De Dombal fut un des premiers à mettre au point et étudier la performance dans la vie réelle d'un tel *système expert* pour un diagnostic médical[10]. Parmi les premiers, Mycin[11] permet l'identification d'infections bactériennes et Sphinx[12] la détection d'ictères, en s'appuyant sur les connaissances dans un domaine précis et une formalisation des raisonnements des experts du domaine.

La deuxième famille d'algorithmes d'IA s'appuie sur l'apprentissage automatique (*machine learning*, ML) à partir d'immenses bases de données résultant pour partie de la pratique médicale humaine et pour partie du fonctionnement des capteurs et d'évolutions objectives de l'état des patients (par exemple, le génome d'un patient est associé à sa durée de survie face à un cancer). Avec le temps, la massification et le partage de ces données, la mise à disposition facile et économique de systèmes de stockage (les *clouds*), l'amélioration des algorithmes de ML et l'exploitation de processeurs de plus en plus puissants (et de plus en plus parallèles), les résultats de ces approches sont devenus de plus en plus convaincants.

Le ML est un sous-domaine de l'IA dans lequel le modèle n'est pas explicitement programmé pour suivre un ensemble d'instructions afin de résoudre une tâche donnée. Il apprend de lui-même comment décider ou

comment agir, en acquérant son propre ensemble de règles à partir des données qu'on lui présente. L'apprentissage profond (*deep learning*, DL) va encore plus loin car il fonctionne en minimisant le recours à l'humain pour nettoyer et prétraiter les données nécessaires à son apprentissage. Cela devient la responsabilité des premières couches des neurones qui les caractérisent. Si ces étapes préalables sont laissées à l'humain cela peut même conduire ces systèmes dans des solutions sous-optimales et il vaut mieux dès lors laisser ces systèmes apprendre et décider d'eux-mêmes ce qu'il y avait de plus « informatif » dans les données. Cette approche est particulièrement adaptée pour les tâches complexes, lorsque toutes les caractéristiques des objets à traiter ne peuvent pas être explicitées ou catégorisées en amont.

En ML, trois grandes familles d'apprentissages peuvent être distinguées : supervisé, non supervisé et par renforcement. **L'apprentissage supervisé** repose sur l'analyse de données labellisées par des humains ou de manière automatique. Le modèle utilise des paires entrée-sortie dans le but d'apprendre la fonction qui corrèle le mieux l'entrée avec la sortie, en sorte qu'elle soit finalement capable de prédire la sortie d'un ensemble de données non labellisées. **L'apprentissage non supervisé** repose sur le traitement de données non labellisées - minimisant le besoin d'interaction humaine, tout en recherchant au mieux à regrouper les données (et ainsi leur associer un label, mais, cette fois, *a posteriori* comme résultat de l'apprentissage). **L'apprentissage par renforcement** fait agir l'agent apprenant dans un environnement complexe qu'il perçoit et sur lequel il agit. Il apprend en réponse aux récompenses ou aux pénalités qui résultent des actions entreprises. Cela permet à l'agent d'apprendre la séquence d'actions qui conduit aux récompenses les plus élevées dans l'accomplissement de sa tâche.

Les réseaux de neurones profonds (*deep neural network*, DNN), un sous-ensemble des réseaux de neurones, eux-mêmes un sous-ensemble des algorithmes de ML, identifient d'eux-mêmes les caractéristiques les plus discriminantes cachées dans les données d'apprentissage. Ces réseaux cherchent au mieux à associer les entrées et les sorties caractérisant les données d'apprentissage. Leur mise au point débute par un entraînement sur un grand nombre d'exemples (des millions voire plus), ce qui leur permet de se forger une *expérience* par la reconnaissance de caractéristiques et d'effectuer des classifications. Ils procèdent par une première phase d'apprentissage et puis de validation afin de sélectionner les architectures neuronales les plus performantes. La dernière phase consiste en un test de leurs capacités sur des données qui n'ont pas été utilisées lors des deux premières phases, permettant ainsi de se faire une idée des performances dont ils seront capables dans l'avenir, face à des situations toutes nouvelles.

L'anamnèse

De nombreux facteurs autres que le langage peuvent intervenir dans une anamnèse bien conduite. La manière de poser les questions du médecin peut susciter des réponses ambiguës, car le patient intègre ses plaintes dans son propre cadre de référence, différent de celui qui l'interroge. La précision des termes utilisés, l'empathie, la tonalité de la voix et même le regard du médecin peuvent influencer considérablement les réponses. Le langage corporel du patient intervient aussi dans la manière qu'un clinicien expérimenté a de procéder.

Obtenir une interaction directe via le langage ainsi que l'enregistrement de toutes ces informations contextuelles et ces signaux subtils reste un exercice compliqué pour un programme informatique. Paradoxalement, les difficultés concernent surtout la communication d'événements banals, en raison des structures grammaticales hiérarchisées et de leur contenu sémantique. En revanche, des termes plus techniques sont plus facilement pris en compte, grâce notamment à l'utilisation de dictionnaires spécialisés. Des progrès notables en *Natural Langage Processing* sont à mettre au crédit d'algorithmes comme BERT et en particulier bioBERT, qui permet le traitement des dossiers médicaux électroniques[13]. Bingli permet une anamnèse guidée avant même la consultation, permettant un gain de temps non négligeable et une proposition de diagnostics[i], à partir de modèles graphiques et de dictionnaires élaborés. D'autres applications existent également comme Babylon[ii], Ada[iii] et Intermedica[iv].

L'examen clinique

Il ne fait guère de doute que la sémiologie demeure un domaine pour lequel l'humain reste supérieur à la machine. Il n'existe pas encore de robots palpeurs ou auscultateurs. Cependant, l'assistance de l'IA peut se révéler très utile, comme en dermatologie[14] ou en ophtalmologie[15]. Les recherches s'égarent parfois dans le futile, comme cet algorithme d'apprentissage profond basé sur des photos du visage, montrant des résultats supérieurs à un score clinique et au modèle de Diamond et Forrester dans la détection de la maladie coronaire[16,17].

Performance des tests diagnostiques et analyse de données

Une différence essentielle entre l'apprentissage humain et l'apprentissage automatique réside dans les associations générales et complexes que les humains parviennent à accomplir à partir de petites quantités de données[18]. En revanche, les techniques de ML présentent les meilleures performances dès qu'il convient de traiter de très grandes quantités de données. Leur utilisation pour l'interprétation et la classification des images est une assistance précieuse en radiologie et en anatomo-pathologie[3]. Un avantage supplémentaire réside dans l'absence de fatigue et autres vécus sensibles ou émotionnels, ainsi que la disponibilité totale, 24 heures par jour, 7 jours sur 7.

(i) http://www.bingli.health/smart-anamnesis/

(ii) https://www.babylonhealth.com/en-gb/regulatory

(iii) https://ada.com/

(iv) https://infermedica.com/product/symptom-checker m

Le domaine cardiovasculaire, évolue vers des modèles prédictifs intégrant des données cliniques et démographiques auxquelles sont associés des paramètres radiologiques comme le score coronaire calcique d'Agatston ou des paramètres scintigraphiques[19,20]. De grands progrès ont été réalisés dans l'interprétation des ECG et en particulier dans le dépistage de la fibrillation auriculaire, mais il reste encore beaucoup à accomplir. Chang *et al.* ont développé un modèle capable à la fois d'identifier les infarctus STEMI et 12 rythmes cardiaques à partir de 60 537 ECG 12 dérivations enregistrés chez 35.981 patients. L'aire sous la courbe ROC (*receiver operating characteristic*) de leur modèle est de 0,987, ce qui est supérieur à celle des cardiologues (0,898), des urgentistes (0,820), des internistes (0,765) et d'un algorithme commercial (0,845)[21].

L'ECG peut être utilisé pour dépister l'insuffisance cardiaque. Attia *et al*. ont utilisé des données appariées d'ECG 12 dérivations et d'échocardiogramme, provenant de 44.959 patients. Ils ont entraîné leur DNN afin d'identifier les patients présentant un dysfonctionnement ventriculaire, défini comme une fraction d'éjection inférieure à 35 %. Testé sur un ensemble indépendant de 52.870 patients, leur modèle a donné des valeurs pour la sensibilité et la spécificité de 86,3 % et 85,7 % respectivement[22]. Un DNN peut même aller jusqu'à prévoir quels individus vont présenter de la FA, même plusieurs années à l'avance à partir de l'ECG de routine[23,24]. Cela pourrait permettre de mettre en évidence des mesures de prévention, telles des adaptations de l'hygiène de vie.

Un groupe de 120 pneumologues a réalisé 6.000 interprétations indépendantes d'épreuves fonctionnelles respiratoires (EFR). Un modèle d'IA a examiné les mêmes données. Les recommandations de l'*American Thoracic Society* et de l'*European Respiratory Society* ont été utilisées comme référence pour l'interprétation. Les pneumologues ont donné une interprétation correcte des EFR dans 74,4 ± 5,9 % des cas. Le logiciel basé sur l'IA a, quant à lui, correctement interprété celles-ci dans 100 % des cas[25].

En néphrologie, un DNN a sélectionné les patients présentant un risque élevé de néphropathie à IgA progressive avec plus de précision que les néphrologues expérimentés (résultats corrects pour 87,0 % des patients avec une sensibilité de 86,4 % et une spécificité de 87,5 %), contre 69,4 % de résultats corrects pour les néphrologues avec une sensibilité de 72 % et une spécificité de 66 %[26].

En gastroentérologie, les techniques d'IA peuvent être utilisées pour faciliter l'analyse des lésions inflammatoires ou des saignements gastro-intestinaux, pour évaluer la fibrose du foie et pour différencier les patients atteints d'un cancer du pancréas de ceux qui souffrent de pancréatite[27]. Deux essais contrôlés randomisés publiés ont comparé la performance de l'endoscopie avec ou sans l'aide d'algorithmes basés sur l'IA. La première étude a testé la capacité d'un système d'IA en temps réel, WISENSE, à surveiller les angles morts pendant la gastroscopie. Au total, 324 patients ont été répartis au hasard pour subir une gastroscopie avec ou sans WISENSE. Le taux d'angles morts était significativement plus faible dans le groupe WISENSE que dans le groupe témoin (5,9 % contre 22,5 %)[28]. La deuxième étude a examiné l'effet d'un système de détection automatique des polypes basé sur DL pendant la coloscopie. Au total, 1.058 patients ont été répartis au hasard dans des groupes qui ont subi une coloscopie diagnostique avec ou sans cette assistance. Le système d'IA a augmenté de façon significative le taux de détection des adénomes, passant de 20,3 % à 29,1 %, et le nombre moyen d'adénomes par patient de 0,31 à 0,53[29].

En utilisant un ensemble de données multicentrique de 28.953 mammographies, McKinney *et al.* ont développé un modèle qui prédit le développement du cancer du sein à 2 ans avec des performances supérieures à celles des radiologues utilisant les critères du *Breast Imaging Reporting And Data System*[30]. Une excellente revue de l'utilisation des techniques d'IA pour le dépistage, le diagnostic et le traitement des lésions cancéreuses a été récemment publiée[31].

L'interprétation des enregistrements électroencéphalographiques est un autre domaine où l'apprentissage automatique a été utilisé. Des algorithmes permettant de détecter les signes d'activité épileptique ont été développés[32,33]. Une combinaison de plusieurs algorithmes de DL a identifié les signatures d'un état cérébral pré-ictal jusqu'à une heure avant une crise[34].

Les traitements

Les différents niveaux de preuve et la diversification nécessaire des entités cliniques a fait exploser le volume des recommandations émanant des experts : par exemple, les dernières *guidelines* de la Société européenne de Cardiologie pour le seul traitement de l'insuffisance cardiaque comportent 128 pages[35]. Ceci complique la tâche des cliniciens de terrain et en particulier des médecins généralistes ; mais sélectionner le traitement optimal pour un patient donné devrait être une tâche envisageable pour un algorithme d'IA dans la mesure où l'application de règles en fonction des scores cliniques peut être facilement implémentée.

Cependant les nombreuses données des dossiers médicaux informatisés (DMI) ne sont malheureusement pas facilement exploitables pour pouvoir développer des modèles plus complexes, mais plus précis et actualisés des risques individuels, par exemple celui de développer une complication cardiovasculaire. Néanmoins, un bel exemple d'intégration au niveau national a été rapporté récemment en Nouvelle-Zélande. Basé sur plus de 400.000 individus, le logiciel PREDICT a mis en évidence une surestimation des calculateurs de risque conventionnels et l'importance de nouvelles données ethniques ou socio-économiques[36].

Les systèmes d'aide à la décision clinique *(Clinical Decision Support Systems, CDSS*) sont des logiciels spécifiquement conçus pour la prise de décision clinique. Ils permettent de mettre en correspondance les caractéristiques d'un patient individuel avec une base de connaissances cliniques informatisée. Ils sont composés de 3 parties : (1) les données et le logiciel (un sys-

tème de règles ou un algorithme de ML) ; (2) le moteur d'inférence (qui applique les règles programmées ou déterminées par l'IA et les structures de données aux données cliniques du patient pour générer une sortie ou une action) ; (3) le mécanisme de communication (le site web, l'application ou l'interface du DMI, avec lequel le médecin interagit) qui présente au médecin le résultat de l'analyse. Les fonctions fournies par les CDSS comprennent les diagnostics, les systèmes d'alarme, la gestion des maladies, les prescriptions des examens et des traitements, le contrôle des médicaments[37]. Une méta-analyse récente réalisée à partir de 122 études randomisées, impliquant 1.203.053 patients et 10.790 prestataires, montre que les CDSS apportent entre 5,8 % et 8,5 % d'amélioration des processus de soins ciblés. Toutefois, les auteurs insistent sur la grande hétérogénéité de ces études et notent que celles menées dans des établissements possédant une longue expérience de l'informatique clinique ont montré des améliorations nettement plus importantes[38].

Un ambitieux projet, *Watson Health*, a été lancé en 2015 par IBM. Watson est un système cognitif dans lequel les composantes du langage, du raisonnement et de l'apprentissage sont regroupées afin de répondre à des questions ou d'explorer de nouveaux liens. Il doit son succès et sa réputation à sa victoire lors d'un jeu télévisé assimilable à « Questions pour un champion ». Dans sa version médicale, il exploite des volumes de données considérables dans différents domaines, comme les images diagnostiques de centaines de millions de patients, les dossiers médicaux des hôpitaux et des assurances, les brevets, la génomique, la chimie, la pharmacologie, les facturations et les publications médicales. Watson a été développé au départ d'une compréhension spécifique de la terminologie scientifique (une ontologie propre à la médecine), ce qui lui permet d'établir de nouveaux liens dans des millions de pages de texte. Watson a été appliqué avec succès à quelques études pilotes dans les domaines de la recherche de nouvelles molécules et de la reformulation de certains composés[39]. La société a ensuite travaillé avec les centres cliniques oncologiques américains les plus réputés dans le but de suggérer aux oncologues les traitements les plus appropriés à chaque patient.

En oncologie toujours, les progrès récents du ML pour prédire le traitement ou la combinaison optimale pour un patient donnent des résultats prometteurs[40]. Un CDSS doté d'une approche probabiliste de ML basée sur des valeurs aberrantes a réussi à générer des alertes cliniquement utiles[41]. Des modèles d'arbres décisionnels se sont révélés très performants pour détecter les prescriptions surdosées ou sous-dosées avec des sensibilités et spécificités de plus de 95 %[42]. Des méthodes d'apprentissage automatique ont été utilisées pour identifier des caractéristiques pertinentes à partir de données génomiques et pour développer des classifications permettant de prédire l'efficacité des médicaments dans les lignées cellulaires cancéreuses[43].

DISCUSSION

Les résultats obtenus par le DL dans le domaine de l'interprétation des actes techniques spécialisés semblent très prometteurs. Il faut cependant attirer l'attention sur la méthodologie prioritairement utilisée dans toutes ces études. Les paramétrages utilisés pour la machine sont réalisés par des experts ou à partir de valeurs de références établies. Ces données labellisées constituent ce que les informaticiens appellent la *ground truth*. C'est sur cette dernière que l'entraînement des algorithmes est réalisé. Le testing du logiciel est le plus souvent réalisé par comparaison avec des praticiens du terrain. C'est un peu comme si une équipe sportive était entrainée avec des professionnels et qu'elle rencontrait ensuite des amateurs. Par ailleurs, la grille d'évaluation correspond à des valeurs établies pour les tests effectués, mais pas nécessairement pour obtenir une amélioration du diagnostic clinique final du patient. Par exemple, Giudicessi *et al.* ont démontré la capacité d'un réseau profond à reconnaître un allongement cliniquement significatif de l'intervalle QT corrigé (QTc › 500 ms) sur des tracés d'appareils ECG mobiles. Ces résultats étaient similaires aux mesures du QTc basées sur l'ECG à 12 dérivations, déterminées à la fois par un cardiologue expert en QT et un laboratoire central commercial[44]. Ceci signifie que l'IA diagnostique l'allongement de l'intervalle QT aussi bien que les experts. Il aurait cependant été plus intéressant de montrer que l'IA pouvait prévoir la survenue des torsades de pointes, une démonstration qui nécessite une toute autre expertise. En effet, toutes ces applications du DL sont prédictives mais fondamentalement non causales (selon le fameux adage « corrélation n'est pas causation »). Elles font « simplement » correspondre des entrées à des sorties, mais ne tiennent pas compte de la causalité. Les tâches d'inférence causale nécessitent les connaissances des experts, non seulement pour spécifier la question, identifier les sources de données pertinentes et éliminer d'éventuels facteurs de confusion, mais aussi décrire la structure causale du système étudié. Aucun algorithme à ce jour ne peut quantifier la précision des inférences causales à partir de données d'observation[45].

Les méthodes statistiques de ML reposent sur l'hypothèse selon laquelle la distribution des échantillons d'apprentissage est représentative de ce qui doit être traité dans la vie réelle, ce qui peut créer des déficiences majeures dans les utilisations qui suivent leur apprentissage. En particulier, les modèles DL sont mis à rude épreuve lorsqu'ils rencontrent des situations peu échantillonnées dans l'ensemble de données d'apprentissage, voire absentes de ces dernières. Les effets de ce que l'on appelle les cygnes noirs - des événements imprévisibles et ayant un impact massif - peuvent être particulièrement préjudiciables lorsqu'on applique un modèle pré-entraîné pour l'inférence. L'utilisation des méthodes de validation croisée peut en partie remédier à ce déséquilibre en le mettant en évidence et en indiquant les voies résolutives (quels types de données font défaut). L'apprentissage à partir de peu de données, ce que les humains réussissent plutôt bien, demeure un défi majeur pour l'IA.

La difficulté d'interprétation des études et leur validation rend encore plus difficile pour l'IA apprenante l'exploitation des publications scientifiques. Malgré le travail assidu des relecteurs et des éditeurs, il peut exister des biais dans la littérature, rendant l'assimilation correcte des résultats par les systèmes d'IA encore plus complexe. Par exemple, Simonato *et al.* rapportent que les taux d'événements dans les bras contrôle étaient souvent surestimés de 28 % en moyenne dans 58 études cardiologiques récentes basées sur la non-infériorité. 95 % de ces études ont utilisé des marges de non-infériorité absolues alors qu'il aurait fallu utiliser des marges relatives, ce qui a conduit à une fourchette de non-infériorité plus permissive que celle proposée initialement[46].

Même l'évolution des techniques statistiques peut compliquer la tâche de l'IA cognitive comme le fait que le sacro-saint seuil de 5 % de signification généralement accepté pour la valeur de p, représentant l'erreur de première espèce (les chances de rejeter l'hypothèse nulle alors qu'elle est vraie) est actuellement remis en question[47,48]. Un système cognitif pourrait aussi être perturbé par le fait que des études observationnelles publient des résultats contradictoires ou que les études rapportant des résultats négatifs récoltent un taux d'acceptation pour publication moindre que les études positives[49], sans vouloir pousser la provocation jusqu'à dire que la plupart des études sont biaisées[50]. Ces nombreux biais concernant les publications scientifiques pourraient permettre d'expliquer au moins en partie l'échec de *Watson Health*, basé pour l'essentiel sur l'exploitation intensive des textes de nature médicale. IBM a fini par abandonner le projet après 7 ans d'efforts, engagé 7.000 personnes et dépensé près de 5 milliards de dollars[51].

L'explicabilité des modèles de DL et l'extraction des caractéristiques utilisées par ces modèles souvent appelés « boîtes noires » font ces dernières années l'objet de travaux très importants. Par ailleurs, utiliser des modèles qui soient interprétables en premier lieu, plutôt que d'essayer de les expliquer *a posteriori* semble une autre alternative[52]. Néanmoins, il ne faut pas rejeter l'utilisation de la boîte noire sous prétexte qu'on ne dispose pas d'explications suffisamment détaillées. Ignorer le détail du fonctionnement des pièces du moteur d'une voiture n'empêche pas de savoir conduire correctement. Une description de la méthodologie reste cruciale, d'autant qu'en ce domaine, les méthodes utilisées peuvent se révéler extrêmement complexes. Une récente revue de 33 études fait état d'un manque d'informations sur les bases de données utilisées qui en limite la reproductibilité d'environ 80 %. Les informations sur les algorithmes d'apprentissage automatique étaient jugées insuffisantes dans 27 % des cas[53]. Des recommandations ont été publiées pour éviter ces problèmes[54].

Tout ceci pourrait permettre d'expliquer pourquoi une majorité de cliniciens pourraient ne pas faire confiance à ces boîtes noires pour lesquelles ils ne peuvent appréhender le raisonnement final menant au diagnostic proposé. En l'absence d'une meilleure compréhension, ils risquent ainsi de douter du résultat, avec pour effet de limiter l'usage de nombreux modèles[55].

Mais ces remarques d'ordre théorique ne doivent surtout pas empêcher la recherche et le développement des applications médicales de l'IA. Il existe en effet de nombreux domaines où ces nouvelles techniques peuvent apporter une valeur ajoutée aussi bien pour le diagnostic, le traitement, la prévention, la prévision et le suivi des maladies. Tout d'abord, le clinicien pourrait utiliser une IA comme assistant personnel, un peu comme s'il disposait en permanence d'un expert à ses côtés pour l'aider dans sa prise de décision. Mieux encore, l'utilisation de l'IA peut permettre de voir ce qui reste invisible à l'œil humain. Par exemple, un DNN peut prévoir quels individus vont présenter de la fibrillation auriculaire jusqu'à plusieurs années à l'avance à partir d'un simple ECG de routine[24]. Dans ce cas, les médecins utilisent l'outil IA comme un super-microscope permettant d'extraire des informations invisibles même aux yeux des meilleurs experts. De plus, la prévision de l'évolution d'une épidémie fait également partie des possibilités de l'IA. Un DNN a généré de bonnes estimations des cas de COVID-19 au Qatar, en Espagne et en Italie. Ces résultats ont montré la grande capacité de généralisation du modèle pour les prédictions à long terme de la pandémie de COVID-19[56]. L'IA peut également être utile pour la prévention des maladies chroniques. Un modèle basé sur des réseaux neuronaux graphiques a présenté une précision de 93,49 % pour la prédiction des maladies cardiovasculaires et de 89,15 % pour celle des maladies pulmonaires chroniques[57]. Tout ceci sans parler de la pléthore d'applications d'IA embarquées sur les objets connectés (les *wearables*) utilisés dans le cadre de la santé mobile - ou « *mHealth* » - qui est elle-même une composante de la santé numérique, définie par l'OMS comme une pratique médicale et de santé publique soutenue par des appareils, tels que les smartphones, les dispositifs de suivi des patients, les assistants numériques personnels et autres systèmes sans fil. Ceux-ci permettent un dépistage et un suivi de différentes pathologies, en impliquant d'avantage les patients tout en dégageant les médecins d'une partie de la charge de travail.

Avant son utilisation, l'IA doit donc être validée par comparaison à des valeurs de références données par des experts : la *ground truth*, en quelque sorte l'équivalent de la vérité objective. La performance prédictive des algorithmes dépend donc de la qualité des annotations faites par ces experts. Ce problème est particulièrement pertinent dans le domaine de l'imagerie, où les erreurs d'annotation et la variabilité inter-observateurs ne sont pas toujours négligeables. Dans un processus typique de labellisation, les différents experts humains fournissent leurs estimations des « vrais » labels sous l'influence de leurs propres biais et niveaux de compétence. Supposons un instant pour fixer les idées que les radiologues experts protocolent les clichés diagnostiquant une tumeur avec 80 % de fiabilité. Dans pareil cas, la *ground truth* demeure chose humaine et il est donc possible d'entraîner un réseau neuronal afin qu'il fasse au moins aussi bien. On retrouve les avantages non pas d'un dépassement mais d'une automatisation de la meilleure compétence

médicale, disponible de manière fiable en tout lieu et à tout moment. Il est toutefois possible qu'une IA, à force d'absorption de données massives et d'entraînement, puisse dépasser ces 80 %, en raison précisément des défaillances cognitives des humains qui les empêchent d'atteindre les 100 % d'exactitude diagnostique (la vérité absolue, la véritable *ground truth*). Voilà la révolution que les neurones de l'IA apprenante sont en train de nous dévoiler. En effet, le spectaculaire succès des systèmes de DL pour des tâches aussi simples que la reconnaissance de chiens ou de chats à partir de photographies a été récemment démontré. On connaît avec certitude l'animal photographié (évidemment sous des angles et des éclairages difficiles) qu'on donne à reconnaître à l'homme et à la machine. La performance humaine la meilleure se situe à 80 % alors que la machine atteint 95 %.

Le même phénomène de supériorité de la machine sur l'homme existe en oncologie[58]. Par exemple, il n'est plus exclu qu'en associant les résultats des examens paracliniques au taux de survie, une IA trouve des corrélations qui auraient échappé à l'humain et oblige ainsi à modifier la classification en vigueur. L'apprentissage non supervisé a ainsi réussi à identifier des sous-groupes de pronostic différent de cancer du poumon et du sein[59,60] et il a permis de découvrir des classifications plus pertinentes pour l'établissement du pronostic à partir des données génétiques[61]. C'est encore une autre manière pour l'IA de surpasser l'homme. Bien que les ayant imitées au départ, l'IA apprenante dépasse les limitations cognitives des praticiens humains.

CONCLUSION

L'utilisation des techniques d'IA afin d'optimiser les prises en charge des patients reste donc un défi. Le ML (machine learning) dépasse l'humain dans l'analyse de grandes quantités de données ce qui explique que ces modèles soient très performants dans l'interprétation de certains tests diagnostiques basés sur un traitement d'images et de signal. Les principaux problèmes à résoudre résident dans la reproductibilité des études, la généralisation des modèles et l'accessibilité à des bases de données fiables. Les techniques d'IA logiques et symboliques correspondent mieux à la manière de raisonner des humains mais leur mise en œuvre reste complexe. Des progrès doivent encore être accomplis par la fertilisation mutuelle de l'IA apprenante et des systèmes basés sur la logique. Il restera encore et toujours à vaincre la résistance naturelle de l'homme à accepter de nouvelles méthodes de diagnostic et de traitement, davantage encore lorsqu'il a le sentiment de n'en pas maîtriser toute la compréhension.

Conflits d'intérêt : néant.

Ce travail a bénéficié du soutien de la Communauté française de Belgique (*FRIA Funding*).

RETROUVEZ TOUTES LES INFORMATIONS CONCERNANT LE NOUVEAU CERTIFICAT INTER-UNIVERSITÉS EN INTELLIGENCE ARTIFICIELLE EN MÉDECINE ET SANTÉ DIGITALE EN PAGE 281

BIBLIOGRAPHIE

1. Artificial Intelligence in Healthcare report | Shaping Europe's digital future [Internet]. [cited 2022 Feb 3]. Available from: https://digital-strategy.ec.europa.eu/en/library/artificial-intelligence-healthcare-report
2. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med. 2020;7:27.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
4. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A *et al*. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-97.
5. Kahneman D. Thinking, fast and slow. Macmillan; 2011.
6. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974;185(4157):1124-31.
7. Klein GA. Sources of power: How people make decisions. MIT press; 2017.
8. Newell A, Simon HA. Computer Simulation of Human Thinking: A theory of problem solving expressed as a computer program permits simulation of thinking processes. Science. 1961;134(3495):2011-7.
9. Bersini H. De l'intelligence humaine à l'intelligence artificielle. Ellipses; 2006.
10. De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J. 1972;2(5804):9-13.
11. Buchanan BG, Shortliffe EH. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project;1984.
12. Fieschi M, Joubert M, Fieschi D, Roux M. SPHINX—A system for computer-aided diagnosis. Methods Inf Med. 1982;21(03):143-8.
13. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):1-13.
14. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
15. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama. 2016;316(22):2402-10.
16. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med. 1979;300(24):1350-8.
17. Lin S, Li Z, Fu B, Chen S, Li X, Wang Y *et al*. Feasibility of using deep learning to detect coronary artery disease based on facial photo. Eur Heart J. 2020;4400-11.
18. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58.
19. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z *et al*. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. Eur Heart J. 2020;41(3):359-67.
20. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S *et al*. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. JACC Cardiovasc Imaging. 2018;11(11):1654-63.
21. Chang K-C, Hsieh P-H, Wu M-Y, Wang Y-C, Wei J-T, Shih ESC *et al*. Usefulness of multi-labelling artificial intelligence in detecting rhythm disorders and acute ST-elevation myocardial infarction on 12-lead electrocardiogram. Eur Hear Journal-Digital Heal. 2021;2(2):299-310.
22. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G *et al*. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med. 2019;25(1):70-4.
23. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ *et al*. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861-7.
24. Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P *et al*. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. Circulation. 2022;145(2):122-33.
25. Topalovic M, Das N, Burgel P-R, Daenen M, Derom E, Haenebalcke C *et al*. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. Eur Respir J. 2019;53(4):1801660.
26. Geddes CC, Fox JG, Allison MEM, Boulton-Jones JM, Simpson K. An artificial neural network can select patients at high risk of developing progressive IgA nephropathy more accurately than experienced nephrologists. Nephrol Dial Transplant. 1998;13(1):67-71.
27. Le Berre C, Sandborn WJ, Aridhi S, Devignes M-D, Fournier L, Smail-Tabbone M *et al*. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology. 2020;158(1):76-94.
28. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J *et al*. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut. 2019;68(12):2161-9.
29. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X *et al*. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813-9.
30. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H *et al*. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
31. Nagy M, Radakovich N, Nazha A. Machine learning in oncology: What should clinicians know? JCO Clin Cancer Informatics. 2020;4:799-810.
32. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput Biol Med. 2018;100:270-8.
33. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng. 2019;16(5):51001.
34. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. IEEE Trans Biomed Circuits Syst. 2019;13(5):804-13.
35. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M *et al*. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contributio. Eur Heart J. 2021;42(36):3599-726.
36. Pylypchuk R, Wells S, Kerr A, Poppe K, Riddell T, Harwood M *et al*. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. Lancet. 2018;391(10133):1897-907.
37. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3(1):1-10.

38. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G *et al.* Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ. 2020;370:m3216.
39. Chen Y, Argentinis JDE, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38(4):688-701.
40. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol. 2020;4(1):1-10.
41. Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Informatics Assoc. 2019;26(12):1560-5.
42. Nagata K, Tsuji T, Suetsugu K, Muraoka K, Watanabe H, Kanaya A *et al.* Detection of overdose and underdose prescriptions—An unsupervised machine learning approach. PLoS One. 2021;16(11):e0260315.
43. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol Cancer Res. 2018;16(2):269-78.
44. Giudicessi JR, Schram M, Bos JM, Galloway CD, Shreibati JB, Johnson PW *et al.* Artificial Intelligence–Enabled Assessment of the Heart Rate Corrected QT Interval Using a Mobile Electrocardiogram Device. Circulation 2021;143:1274-86.
45. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. https://doi.org/101080/09332480201915795 78 [Internet]. 2019 Jan 2 [cited 2022 Feb 20];32(1):42-9.
46. Simonato M, Ben-Yehuda O, Vincent F, Zhang Z, Redfors B. Consequences of Inaccurate Assumptions in Coronary Stent Noninferiority Trials: A Systematic Review and Meta-analysis. JAMA Cardiol. 2022;7(3):320-7.
47. Nuzzo R. Scientific method: statistical errors. Nat News. 2014;506(7487):150.
48. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Vol. 70, The American Statistician. Taylor & Francis; 2016:129-33.
49. Nimpf S, Keays DA. Why (and how) we should publish negative data. EMBO Rep [Internet]. 2020 Jan 7 [cited 2022 Feb 20];21(1). Available from: /pmc/articles/PMC6945059/
50. Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Med [Internet]. 2005 Jul 23 [cited 2022 Feb 20];2(8):e124. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
51. What Ever Happened to IBM's Watson? - The New York Times [Internet]. [cited 2022 Jan 18]. Available from: https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html#commentsContainer
52. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.
53. Olorisade BK, Brereton P, Andras P. Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. J Biomed Inform. 2017;73:1-13.
54. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Harrow J *et al.* DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7.
55. Castelvecchi D. Can we open the black box of AI? Nat News. 2016;538(7623):20.
56. Shawaqfah M, Almomani F. Forecast of the outbreak of COVID-19 using artificial neural network: Case study Qatar, Spain, and Italy. Results Phys. 2021;27:104484.
57. Lu H, Uddin S. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. Sci Rep. 2021;11(1):1-12.
58. Bertsimas D, Wiberg H. Machine Learning in Oncology: Methods, Applications, and Challenges. JCO Clin Cancer Informatics. 2020;4:885-94.
59. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. J Biomed Biotechnol. 2009;2009:632786.
60. Aure MR, Vitelli V, Jernström S, Kumar S, Krohn M, Due EU *et al.* Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Res. 2017;19(1):1-18.
61. Kakushadze Z, Yu W. * K-means and cluster models for cancer signatures. Biomol Detect Quantif. 2017;13:7-31.

Travail reçu le 1er février 2022 ; accepté dans sa version définitive le 5 avril 2022.

CORRESPONDANCE :

J.-M. GREGOIRE
IRIDIA – Université libre de Bruxelles
Av. F. Roosevelt 50 / CP 161 - 1050 Bruxelles
E-mail : jean-marie.gregoire@ulb.be