Avertissement: notes prises au vol… erreurs possibles… prudence!

HUG: Hôpital cantonal de Genève mardi 26 Juillet 2022

Evidence clinique & Intelligence Artificielle: quels enjeux pour la pratique?

Pr Th. Agoritsas

L'intelligence artificielle a pris son essor ces dernières années. Ce qui nous intéresse aujourd'hui, c'est l'impact de l'IA sur la production de preuves et sur la pratique clinique.

Nous vivons actuellement une crise de confiance en la science, qui a été décuplée par le covid…Crise qui s'enfonce avec les données frauduleuses produites par Surgisphere et publiée dans Lancet Gate (hydroxychloroquine et macrolides pour COVID19) ou encore avec la suspicion de fraude sur 20 ans de recherche sur la maladie d'Alzheimer…

Un effort de rigueur concernant la production de connaissance est donc nécessaire.

Un article récent utilise l'IA pour la détection précoce du sepsis, celui-ci est immédiatement repris par les réseaux sociaux comme étant un changement de pratique à mettre en place immédiatement…Néanmoins, c'est un peu rapide, et il faut évaluer l'outil avant tout!

Actuellement, l'IA se déploie surtout dans l'assistance au diagnostic, dans la prédiction du risque et dans la prédiction de la réponse au traitement.

Modèles de prédiction

Il existe plusieurs types de modèles de prédiction. Le modèle statistique traditionnel est le plus utilisé: il donne des prédicteurs qui peuvent être combinés (régression cox, régression logistique…). Pour cela, chaque partie de l'équation doit être définie.

Un autre modèle est l'apprentissage automatique ou "machine learning", qui, à travers des réseaux neuronaux et des bases de données, peut donner des prédicteurs.

Les modèles d'apprentissage profond ou "deep learning" permettent l'analyse d'imagerie et d'apprendre, pixel après pixel, des modèles complexes qui peuvent lire une image.
Et cela, sans aucune injonction humaine, contrairement au machine learning, qui peut être supervisé, modifié et influencé par les chercheurs…

Impact du *machine learning* sur la pratique clinique

Trois voies semblent avoir le plus d'impact: celle de la précision, avec des modèles de prédiction plus performants, celle d'une meilleure identification des sous-groupes de patients, et celle de l'impact clinique direct.

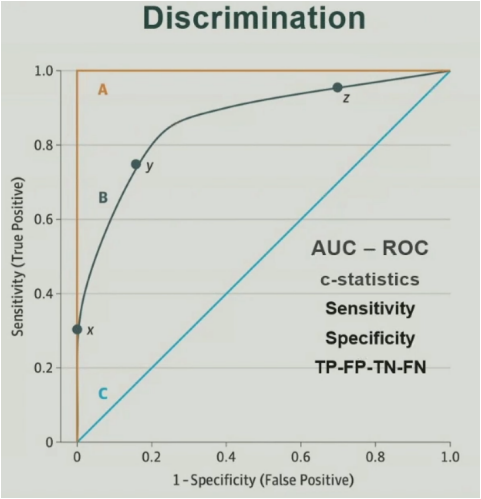
La précision des modèles de prédiction

Une grande précision de modèle ne veut pas dire meilleur résultat. Les modèles de prédiction sont souvent ignorés en clinique, ne sont pas forcément meilleurs que l'instinct clinique, pourraient ne pas changer la prise en charge voire la modifier sans changer le résultat…

Ce guide d'utilisateur , écrit en partie par l'orateur, donne une idée d'évaluation des outils pronostiques. Il est rapidement résumé ci-dessous.

Dans un modèle, il faut scruter deux aspects:

La **discrimination**: faire la différence entre les malades et les non-malades (diagnostic) ou entre les haut et bas risques (pronostic) → sensibilité, spécificité, faux négatifs….et puis il y a la courbe ROC



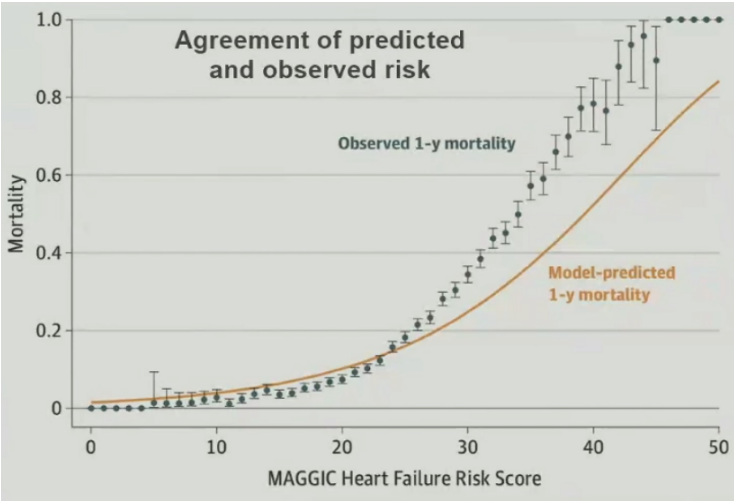
Toutes les performances pronostiques à travers différents strates pronostiques sont représentées par la courbe.

Ici, l'aire sous la courbe AUC est de 80% ou 0.8, ce qui signifie que le modèle va classer le patient correctement 80% du temps. (Le patient à haut risque est annoncé comme à haut risque) …Ce qui veut dire qu'il se trompe 20% du temps.

Essentiel mais pas suffisant pour informer une décision…pour cela il faut évaluer la **calibration**:

A quel point ce que prédit le modèle (ex: 90% de risque d'être malade), correspond à ce que l'on trouve dans la réalité?

Ici on voit que la calibration est bonne dans les scores bas (patients à bas risque) mais mauvaise dans les scores élevés, avec une prédiction inférieure à ce qui est observée.



Nous sommes plutôt habitués à de mauvaises AUC, comme pour le CHA2DS2-VASc, score de prédiction du risque de thrombose, qui est à 0.64…

Un modèle est d'abord dérivé sur un groupe test, puis validé sur un autre groupe de patients avec un contexte différent de préférence.

Donc, sur la voie de la précision, une étude qui présente une discrimination et calibration satisfaisante, et testée sur différentes populations (validée), peut encore être à risque de biais:

Il faut donc utiliser les critères de pronostic GRADE pour établir une certitude sur la discrimination et la calibration du modèle.

Par exemple, cette revue systématique sur 62 publications s'intéresse aux risques de biais dans les modèles de *machine learning* en oncologie, et montre que 84% des modèles développés et 51% des modèles validés sont à haut risque de biais.

Les effets de sous-groupe

Les modèles cherchent à prédire les groupes de patients qui vont le mieux répondre au traitement ou à la prédiction.

Pour rappel, la différence absolue dans la réduction relative du risque n'est pas un effet de sous-groupe, mais bien la variation de gain d'un traitement donné entre des patients qui sont peu à risque et des patients à haut risque de mortalité.

Par exemple, si le traitement baisse la mortalité avec une réduction relative de 50% (RR=0.5), et que le risque basal de mourir est de 20%, il y a un gain de 10% de baisse du risque, alors que si le risque de mourir est de 2%, le gain n'est que de 1%. C'est la différence absolue.

Ce n'est pas un effet de sous-groupe car tous les patients réagissent de la même manière.

La différence relative, elle, pourrait cacher un effet de sous-groupe, avec des populations qui réagissent différemment au traitement et ont des réduction relatives différentes.

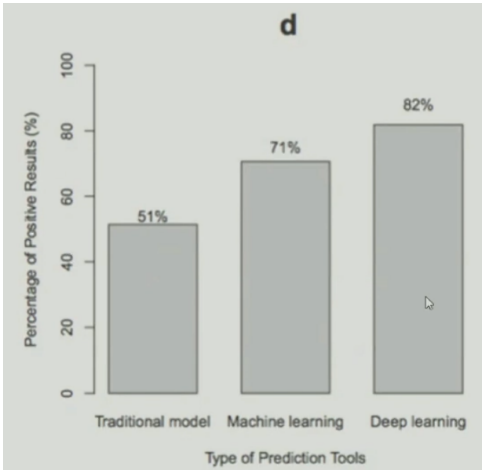
Certaines études, comme celle sur le remdesivir contre le covid, prétendent à un effet de sous-groupe fourni par la boîte noire du machine learning. Cela fait douter de la crédibilité de ces effets, qui pourraient aussi bien être dûs au hasard…

Un autre guide d'utilisateur est disponible pour l'analyse des sous-groupes: les statistiques sont-elles compatibles avec du hasard? La réaction du sous-groupe fait-elle partie des hypothèses? … Pour une analyse plus approfondie, il existe l'outil ICEMAN.

La preuve de l'impact direct en clinique

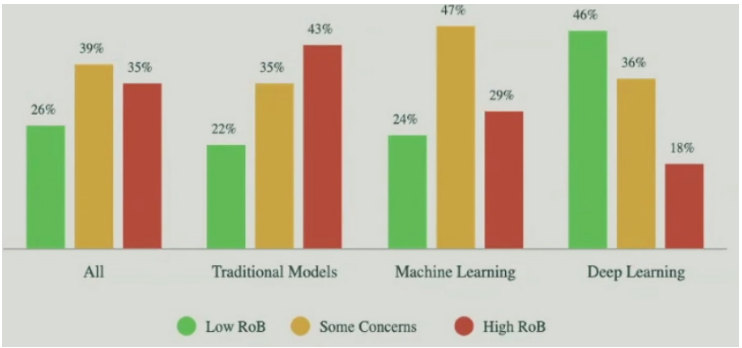
Une revue systématique revoit 65 essais randomisés sur les outils de prédictions. Elle comprend les trois types de modèles statistiques: traditionnel, machine learning et deep learning. Seuls 28 essais sont appuyés sur l'IA.

Les thèmes abordés dans ces études sont: adénomes et détection de polypes, prédiction diagnostique, de durée de séjour, de durée de délais thérapeutique pour les anticoagulants…

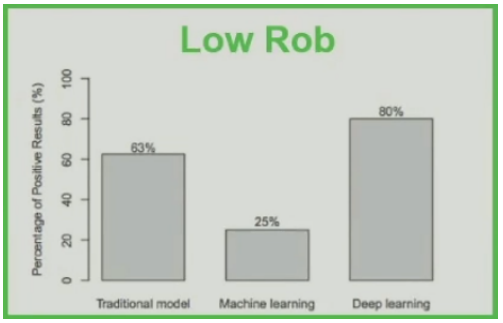


Selon la revue, l'intelligence artificielle a plus de positivité dans les essais randomisés pour montrer l'efficacité…

…mais les risques de biais (RoB) sont importants!



Seules 26% des études sont à bas risque de biais, et les études qui prétendent à 100% de positivité sont à risque de biais élevé.



En prenant uniquement les études à bas risque, le graphique est modifié comme suit:

Cela montre un contraste entre le machine learning et le deep learning.

Le groupe CONSORT fournit des lignes directrices sur la façon de rapporter les essais randomisés, et les a récemment mises à jour pour inclure l'IA: 14 nouveaux items y sont spécifiés.

L'IA doit être mentionnée dans le titre et le résumé, il faut mentionner quelles sont les données entrées dans le modèle et de quels patients elles proviennent, discuter des effets secondaires, du financement et des conflits d'intérêts…

Cas illustratif - dépistage ECG

La dysfonction du ventricule gauche asymptomatique est associée à une baisse de la qualité de vie et à plus de mortalité; il y a donc un intérêt à dépister certains patients.

Cet essai randomisé analyse l'utilisation d'un ECG facilité par IA pour détecter les patients avec une fraction d'éjection abaissée. Elle inclut des patients en ambulatoire, sans symptômes attribuables à la décompensation cardiaque (a priori).

L'intervention est l'assistance par IA de la lecture des ECG, et le contrôle est sans assistance.

Les critères de jugement qui nous intéressent sont la mortalité, les symptômes, les hospitalisations pour décompensation cardiaque, les coûts mais aussi l'anxiété lié au dépistage et les faux positifs/négatifs.

Le modèle est élaboré sur 50'000 patients avec un AUC de 0.93, puis validé sur presque 4000 patients avec la même performance (AUC 0.9!). Mais ces personnes proviennent de la même clinique, il y a peu de variation de population dans la validation.

L'étude comprend 120 équipes de premier recours, 60 sont assistées par IA et 60 pas. Le critère de jugement principal est la baisse de la fraction d'éjection <50%.

Certains détails sont imparfaits et à risque de biais: pas de mention d'aveugle, de comité d'éthique…et 15% des patients sont symptomatiques!

6% des ECG ressortent positifs dans les deux groupes. Dans le groupe assisté par IA, il y a 11,5% de plus d'échographies à la suite de l'ECG (38% vs 50%).

Parmi les échographies réalisées, on note une augmentation de 1.1% de positivité, soit une augmentation de diagnostic chez 0.3% des patients (0.6% à 0.9%). L'effet est petit…

Une modélisation de la mortalité à 5 ans sans diagnostic est faite dans chaque groupe: on passe de 3080 décès à 3072, soit une réduction de 8/100 000 avec le modèle d'IA.

Comment passer de la preuve à la décision? - que décider?

- Signal bénéfices + certitude: petit bénéfice
- Signal de risque + certitude: aucune information
- Valeurs et préférences: probablement différentes selon le patient
- Ressources, Accessibilité, Faisabilité: changement de pratique….faire des ECG à tous
- Equité, justice distributive…

Globalement… l'implémentation globale n'est probablement pas à recommander…

Plus l'IA se développe, plus il faut être intelligent…

Compte-rendu de Valentine Borcic
valentine.borcic@gmail.com
Transmis par le laboratoire MGD
colloque@labomgd.ch