Analyse

Problèmes associés au déploiement des modèles fondés sur l'apprentissage machine en santé

Joseph Paul Cohen PhD, Tianshi Cao MSc, Joseph D. Viviano MSc, Chin-Wei Huang MSc, Michael Fralick MD PhD, Marzyeh Ghassemi PhD, Muhammad Mamdani MSP PharmD, Russell Greiner PhD, Yoshua Bengio PhD

■ Citation : *CMAJ* 2021 September 7;193:E1390. doi : 10.1503/cmaj.202066-f; diffusion hâtive le 30 août 2021

Voir la version anglaise de l'article ici : www.cmaj.ca/lookup/doi/10.1503/cmaj.202066; voir les articles connexes ici : www.cmaj.ca/lookup/doi/10.1503/cmaj.202434-f et www.cmaj.ca/lookup/doi/10.1503/cmaj.210036-f

Dans un article connexe, Verma et ses collègues s'intéressent à la manière dont des solutions fondées sur l'apprentissage machine peuvent être élaborées et mises en place pour appuyer la prise de décision médicale[1]. Les systèmes d'aide à la prise de décision et les outils de prédiction clinique élaborés à l'aide de l'apprentissage machine (y compris le cas particulier de l'apprentissage en profondeur) ressemblent aux outils cliniques créés à partir de modèles statistiques traditionnels; ils ont donc des limites semblables[2,3]. Si un modèle génère des prédictions erronées, ses utilisateurs pourraient commettre des erreurs qu'ils auraient autrement évitées lors des soins aux patients; il est donc important de comprendre les failles de ces modèles[4]. Nous discutons de ces limites — en nous concentrant sur 2 enjeux précis : la généralisation hors échantillon et l'attribution incorrecte des caractéristiques — pour souligner l'importance de connaître les possibles failles des solutions fondées sur l'apprentissage machine.

Points clés

- Les systèmes d'aide à la prise de décision et les outils de prédiction clinique fondés sur l'apprentissage machine (y compris le cas particulier de l'apprentissage en profondeur) ressemblent aux outils cliniques d'aide utilisant des modèles statistiques classiques, et ont donc des limites semblables.
- Si un modèle d'apprentissage machine est formé à l'aide de données ne correspondant pas aux données réelles rencontrées après son déploiement, son rendement pourrait être moins bon que prévu.
- Pendant leur formation, les algorithmes d'apprentissage machine empruntent la voie de la moindre résistance; ils peuvent donc apprendre des caractéristiques faussement corrélées aux résultats attendus plutôt que les bonnes caractéristiques; cela peut nuire à la généralisation efficace par le modèle ainsi généré.
- Pour éviter les erreurs découlant de ces problèmes, il faut évaluer attentivement les modèles d'apprentissage machine à l'aide de nouvelles données réelles et d'échantillons de données qui devraient mettre le modèle à l'épreuve, comme les données sur différentes tranches de la population, des conditions difficiles ou des intrants de mauvaise qualité.

Quelles sont les caractéristiques des modèles fondés sur l'apprentissage machine?

Dans le présent document, le terme « modèle fondé sur l'apprentissage machine » fait référence à un modèle qui a été créé par l'exécution supervisée d'un algorithme d'apprentissage machine sur un ensemble de données étiquetées. Ces modèles sont formés sur des ensembles de données prédéterminées : les données de formation. Il s'agit de données démographiques ou de données provenant de pays, d'hôpitaux, d'appareils, de protocoles, etc. Les modèles fondés sur l'apprentissage machine ne sont pas dynamiques, à moins d'avoir été conçus ainsi — ils ne changent donc pas au fil de leur utilisation. Ils sont typiquement déterministes : ils ont appris un ensemble fixe de poids (coefficients ou paramètres) qui ne change pas lorsque le modèle est exécuté — pour un intrant donné, le modèle donnera toujours la même prédiction. Bien qu'il existe des systèmes adaptatifs, qui peuvent « apprendre » pendant leur utilisation en intégrant de nouvelles données, ces systèmes peuvent donner des prédictions différentes pour un même intrant, et leur sûreté et leur besoin de supervision demeurent incertaines[5].

Nous appellerons « données réelles » les données qu'un modèle fondé sur l'apprentissage machine rencontrera une fois qu'il sera déployé. Si les données de formation d'un modèle ne correspondent pas aux données réelles qu'il rencontrera, le rendement de celui-ci pourrait être moins bon que prévu[6,7] — un problème courant est la généralisation hors échantillon (sujet traité en détail plus loin). Un autre problème peut être que les données de formation contiennent des caractéristiques faussement corrélées aux résultats que l'outil doit prédire; dans ce cas, le modèle peut faire ses prédictions à partir des « mauvaises » caractéristiques (aussi traité plus loin). Le créateur d'un modèle doit donc trouver un ensemble de données de formation qui correspond le plus possible aux données réelles, et les professionnels de la santé qui utilisent l'outil doivent connaître précisément les failles potentielles du modèle et les limites de ses données de formation.

Quels sont les problèmes potentiels des modèles fondés sur l'apprentissage machine?

Généralisation hors échantillon

Les médecins qui viennent d'obtenir leur diplôme sont habituellement plus à l'aise lorsqu'ils traitent des patients présentant des troubles de santé qu'ils ont vus pendant leur résidence, mais ils sont aussi capables de traiter d'autres troubles en utilisant leurs connaissances théoriques pour reconnaître le portrait clinique de différentes maladies. Les méthodes d'apprentissage machine, quant à elles, se limitent aux données fournies durant la phase de développement et de formation. De plus, les modèles fondés sur l'apprentissage machine ne connaissent habituellement pas leurs propres limites, à moins que des éléments soient inclus pour aider le modèle à reconnaître les données qui ne font pas partie de son échantillon (par exemple, il est possible d'intégrer une fonction empêchant un système de diagnostic de radiographies pulmonaires de poser un diagnostic de pneumonie après avoir analysé une photo de chat[8] — voir les stratégies décrites plus loin). Il existe 3 catégories de données hors échantillon[9], qui sont résumées à la figure 1 :

- Les données qui n'ont pas de lien avec la tâche, comme des images visiblement destinées à une autre fin; p. ex., des images de résonnance magnétique présentées à un modèle entraîné pour des radiographies; et des images destinées à une autre fin, mais de manière moins évidente, comme une radiographie du poignet traitée par un modèle entraîné pour les radiographies pulmonaires.
- Les données mal préparées; p. ex., des radiographies pulmonaires floues, ayant un mauvais contraste ou un mauvais positionnement, des images présentées dans le mauvais format de fichier ou incorrectement traitées, et des images provenant du mauvais protocole d'imagerie.
- Les données ne faisant pas partie des données de formation en raison d'un biais de sélection; p. ex., des images montrant une maladie absente des données de formation ou provenant d'une autre tranche de la population que celle utilisée pour les données de formation.

Dans ces cas, un modèle fondé sur l'apprentissage machine aura un rendement sous-optimal ou donnera des résultats inattendus.

De nombreuses stratégies ont été mises au point pour détecter les données hors échantillon et en prévenir le traitement. Une approche courante est de demander au modèle de déterminer à quel point un échantillon correspond à ses données de formation, correspondance qui peut être donnée sous la forme d'une note. Si la note dépasse un certain seuil, le modèle peut choisir de ne pas traiter un échantillon de données. Dans les cas d'interprétation d'images, le modèle tente par exemple de reconstruire l'image et de comparer cette reconstruction à l'image originale en utilisant une mesure de similitude, comme la différence absolue entre les pixels[8,10]. Typiquement, un modèle aura de la difficulté à reconstruire une image ne faisant pas partie de ses données de formation. Si l'image reconstruite est assez similaire pour être jugée « correcte », le modèle peut traiter l'image originale, sinon, il ne la traite pas. Cependant, pour développer et évaluer ces systèmes de détection des données hors échantillon, il faut utiliser des exemples connus de données hors échantillon. Bref, même les stratégies de prévention des erreurs comportent des limites.

Attribution incorrecte des caractéristiques

Les modèles fondés sur l'apprentissage machine utilisent habituellement les ensembles de caractéristiques les moins complexes permettant de différencier les résultats attendus selon l'ensemble de données de formation. Le modèle adopte la voie

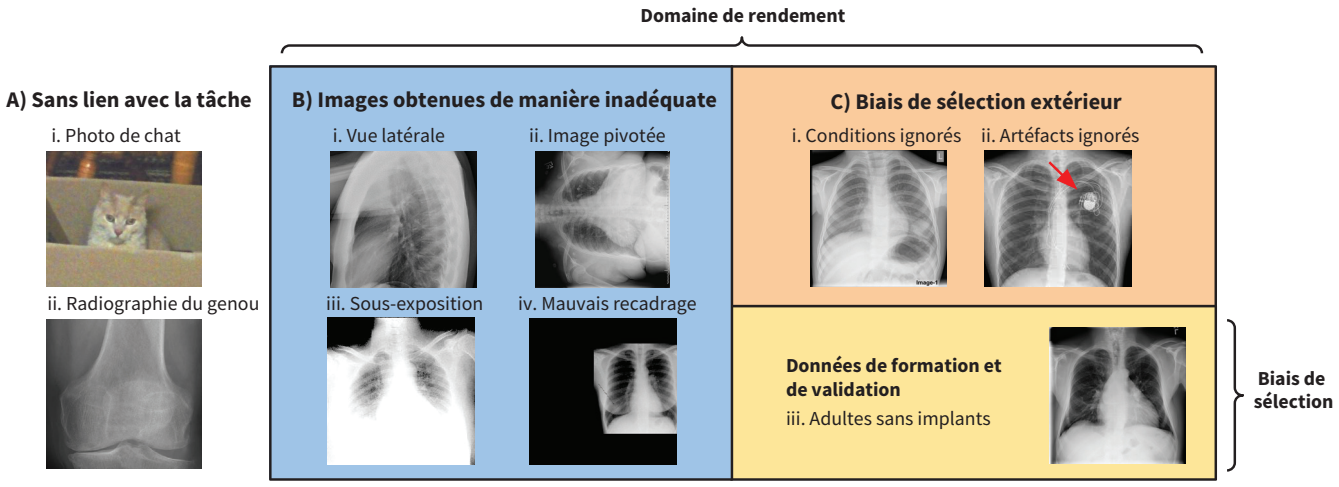

Figure 1 : Cette figure montre 3 catégories de données hors échantillon, toutes dans le contexte de la formation d'un algorithme d'apprentissage machine apprenant à lire les radiographies pulmonaires chez l'adulte (voir l'image C iii). A) Images sans lien avec la tâche. B) Images obtenues de manière inadéquate. C) Éléments visuels ignorés en raison de biais de sélection dans les données de formation (les images présentant des lésions de cancer du poumon et un stimulateur cardiaque n'ont pas été incluses dans les données de formation, et n'ont donc jamais été vues pendant la formation). C) iii) Données de formation sujettes à un biais de sélection.

de la moindre résistance pendant sa formation[11–13], trouvant des caractéristiques qui sont de forts prédicteurs du résultat attendu, ce qui permet de rendre la prédiction juste. Un modèle peut toutefois aussi trouver des distracteurs dans les données qui sont faussement corrélés au résultat attendu[14] et, une fois que cela se produit, le modèle peut cesser de chercher de nouvelles caractéristiques vraiment distinctives, même si elles sont présentes[15]. Par exemple, prenons un modèle qui apprend à lire les radiographies pulmonaires; les distracteurs peuvent être l'hôpital d'origine, les paramètres d'imagerie, la vue de la radiographie (antéro-postérieure ou antéro-postérieure en position couchée), et la présence d'artéfacts, comme un stimulateur cardiaque ou un tube endotrachéal. Si les protocoles cliniques ou le traitement des images changent au fil du temps, le modèle peut détecter des tendances dans les données de formation, qui peuvent agir comme des distracteurs[16]. Ou si des images provenant de différents hôpitaux sont regroupées et que la prévalence d'une maladie varie d'un hôpital à l'autre, un modèle peut apprendre à reconnaître l'hôpital d'origine à partir d'éléments visuels subtils, puis baser ses prédictions sur l'hôpital associé à l'image plutôt que sur l'image elle-même. Cela peut faire en sorte qu'un modèle semble plus juste qu'il ne l'est réellement si les données de test contiennent les mêmes artéfacts (la même distribution par hôpital), mais le même modèle pourrait échouer de manière spectaculaire si les données réelles ne présentent pas ces artéfacts. De plus, les caractéristiques démographiques des patients (âge ou sexe) peuvent être déduites de certaines parties des données de formation et utilisées par un modèle pour prédire la prévalence des résultats (la probabilité antérieure) dans l'ensemble de données de formation si de meilleures caractéristiques liées au résultat d'intérêt sont moins évidentes dans les données.

Les ensembles de données médicales sont souvent relativement petits, ce qui pourrait faire augmenter la probabilité de caractéristiques faussement corrélées au résultat. Des recherches sont en cours pour déterminer comment modifier l'algorithme d'apprentissage des modèles pour éviter ce problème[11,17]. On sait déjà que l'utilisation d'un important ensemble de données diversifiées pour la formation d'un modèle fondé sur l'apprentissage machine aide à atténuer l'effet des distracteurs. D'autres solutions comprennent l'apprentissage non supervisé et l'apprentissage par transfert[18], des processus qui utilisent des données non étiquetées ou des données étiquetées destinées à une autre tâche afin d'entraîner les modèles et d'éviter la détection de caractéristiques faussement corrélées au résultat dans un ensemble précis de données. Ces méthodes permettent habituellement l'utilisation par les modèles d'un volume supérieur de données et accroissent la probabilité qu'un modèle apprenne des caractéristiques qui sont assez générales et utiles pour la tâche visée[18].

Lorsque des caractéristiques propres à pathologie ne sont pas des prédicteurs assez forts dans certaines images, le modèle pourrait être contraint de deviner et de prédire la prévalence d'une maladie ou le résultat selon l'ensemble de données de formation. Dans ce cas, le modèle semblera fonctionner lorsqu'il est appliqué à des données pour lesquelles la prévalence de la maladie est la même que dans les données de formation; il pourrait donner la « bonne » réponse. Toutefois, s'il est appliqué à une autre population où l'on observe une prévalence différente, les prédictions du modèle seront probablement erronnées[19,20], et pourraient entraîner des préjudices. Il est donc important que les créateurs et les utilisateurs des modèles vérifient qu'ils détectent bien des caractéristiques qui sont réellement des prédicteurs du résultat d'intérêt, à l'aide d'une méthode d'attribution des caractéristiques comme la méthode de la descente de gradient[21] ou la création d'un intrant contrefactuel montrant ce qui pourrait changer la capacité de prédiction de l'élément[22] pendant le développement du modèle et après son déploiement.

En lien avec cet élément, une autre préoccupation est que certains modèles pourraient simplement apprendre à imiter les actions des professionnels de la santé lors de la génération des données. Par exemple, si un modèle est entraîné à prédire le besoin d'une transfusion sanguine selon des données historiques sur les transfusions, il pourrait ne trouver dans l'ensemble de données aucun élément informatif qui puisse l'aider dans ses prédictions, et le modèle apprendrait à répliquer les pratiques existantes. Le modèle apprendra alors les « mauvaises habitudes », à moins que l'ensemble de données utilisé pour le créer soit corrigé. Une approche pour remédier à ce problème serait de demander à des experts d'étiqueter l'ensemble de données avec les résultats d'intérêt réels (transfusion appropriée et transfusion inappropriée), mais cette technique pourrait demander beaucoup de ressources, et les experts pourraient ne pas toujours s'entendre sur les étiquettes. Il serait encore mieux d'utiliser seulement des étiquettes objectives qui ne dépendent pas de l'expertise humaine.

Qu'est-ce qui peut atténuer ces problèmes?

Pour éviter les erreurs liées aux problèmes présentés ci-dessus, il faut évaluer attentivement les modèles fondés sur l'apprentissage machine[23] à l'aide de nouvelles données provenant des données réelles, y compris des échantillons qui devraient faire ressortir les failles du modèle, comme celles présentant des caractéristiques démographiques différentes, des conditions difficiles, des images de faible qualité ou des erreurs. Une approche possiblement utile est de créer des ensembles de données de test en équilibrant les données selon des caractéristiques n'ayant aucun lien avec la tâche pour observer les différences dans le rendement du modèle selon des facteurs comme des classes démographiques minoritaires[24] ou des régions géographiques[25]. Prenons par exemple un modèle qui a appris à se concentrer sur une caractéristique faussement associée au résultat, comme l'âge. Le déploiement de ce modèle dans un ensemble de données où l'âge est toujours le même, avec une répartition équilibrée des autres caractéristiques, mènerait à un mauvais rendement. Les résultats de tels tests du rendement d'un modèle devraient être présentés de manière transparente pour montrer ses limites[26]. Un article connexe traite de l'évaluation des modèles fondés sur l'apprentissage machine plus en profondeur[27].

Conclusion

Il faut connaître ces problèmes des modèles d'apprentissage machine et y remédier avant leur déploiement pour que d'importants investissements ne résultent pas en des échecs, qui pourraient être coûteux ou catastrophiques. Par exemple, le programme « Watson for Oncology » d'IBM[28] a été suspendu après un investissement de 62 millions de dollars, supposément en raison de recommandations cliniques problématiques qui ont mené à une faible acceptabilité de l'outil chez les professionnels de la santé. Quant à l'initiative d'apprentissage machine de Google pour détecter la rétinopathie diabétique[29], elle a connu des difficultés lorsqu'elle a été appliquée à de vraies images de moins bonne qualité que les données de formation provenant de cliniques de la Thaïlande, ce qui a causé beaucoup de frustration chez les patients et le personnel. L'anticipation et l'atténuation des problèmes présentés ici seront donc essentielles pour éviter ces échecs coûteux.

Références

1. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ* 2021;193:E1351-7.
2. Liu Y, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806-16.
3. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *AJR Am J Roentgenol* 2019;212:513-9.
4. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020;26:1229-34.
5. Artificial intelligence and machine learning in software as a medical device. Silver Spring (MD): US Food and Drug Administration; 2021. Accessible ici : https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (consulté le 14 mai 2021).
6. Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. *Learning from data: a short course.* AMLbook.com; 2012.
7. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178:1544-7.
8. Cohen JP, Bertin P, Frappier V. Chester: a web delivered locally computed chest x-ray disease prediction system. *ArXiv* 2020 Feb. 2. Accessible ici : https://arxiv.org/abs/1901.11210 (consulté le 14 mai 2021).
9. Cao T, Huang C, Hui DY-T, et al. A benchmark of medical out of distribution detection. uncertainty & robustness in deep learning workshop at ICML. *ArXiv* 2020 Aug 5. Accessible ici : https://arxiv.org/abs/2007.04250 (consulté le 14 mai 2021).
10. Shafaei A, Schmidt M, Little J. A less biased evaluation of out of distribution sample detectors. *ArXiv* 2019 Aug. 20 Accessible ici : https://arxiv.org/abs/1809.04729 (consulté le 16 août 2021).
11. Ross AS, Hughes MC, Doshi-Velez F. Right for the right reasons: training differentiable models by constraining their explanations. *ArXiv* 2017 May 25. Accessible ici : https://arxiv.org/abs/1703.03717 (consulté le 16 août 2021).
12. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018;15:e1002683.
13. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019;2:31.
14. Viviano JD, Simpson B, Dutil F, et al. Saliency is a possible red herring when diagnosing poor generalization. *ArXiv* 2021 Feb. 10. Accessible ici : https://arxiv.org/abs/1910.00199 (consulté le 16 août 2021).
15. Reed RD, Marks RJ. Neural smithing: supervised learning in feedforward artificial neural networks. Cambridge (MA): MIT Press; 1999.
16. Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *ArXiv* 2019 Aug. 2. Accessible ici : https://arxiv.org/abs/1908.00690 (consulté le 16 août 2021).
17. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: *Proceedings of the 32nd International Conference on Machine Learning*; 2015 July 6–11; Lille [FR]. *PMLR* 2015;37:1180-9. Accessible ici : http://jmlr.org/proceedings/papers/v37/ganin15.html (consulté le 14 mai 2021).
18. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*; 2012 Jun. 28–July 2. Bellevue, Wash. [USA]. *PMLR* 2012;27:17-36. Accessible ici : http://proceedings.mlr.press/v27/bengio12a.html (consulté le 14 mai 2021).
19. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, et al. A unifying view on dataset shift in classification. *Pattern Recognit* 2012;45:521-30.
20. Brown MRG, Sidhu GS, Greiner R, et al. ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front Syst Neurosci* 2012;6:69.
21. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *ArXiv* 2014 Apr. 19. Accessible ici : https://arxiv.org/abs/1312.6034
22. Cohen JP, Brooks R, En S, et al. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. *ArXiv* 2021 Apr. 24. Accessible ici : https://arxiv.org/abs/2102.09475 (consulté le 16 août 2021).
23. Powers DMW. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *ArXiv* 2020 Oct. 11. Accessible ici : https://arXiv:2010.16061 (consulté le 16 août 2021).
24. Seyyed-Kalantari L, Laleh G, McDermott M, et al. CheXclusion: fairness gaps in deep chest x-ray classifiers. *ArXiv* 2020 Oct. 16. Accessible ici : https://arxiv.org/abs/2003.00827 (consulté le 16 août 2021).
25. Shankar S, Halpern Y, Breck E, et al. No classification without representation: assessing geodiversity issues in open data sets for the developing world. *ArXiv* 2017 Nov. 22. Accessible ici : https://arxiv.org/abs/1711.08536 (consulté le 16 août 2021).
26. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *FAT* 2019 — Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*; 2019 Jan. 29–31; Atlanta [GA]. Accessible ici : https://doi.org/10.1145/3287560.3287596 (consulté le 14 mai 2021).
27. Antoniou T, Mamdani M. Evaluation of machine learning solutions in medicine. *CMAJ* 2021 Aug. 30 [cyberpublication avant impression]. doi:10.1503/cmaj.210036.
28. Blier N. Stories of AI failure and how to avoid similar AI fails [blog]. Amherst (MA): Lexalytics; 2020 Jan. 30. Available: https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020 (consulté le 18 mai 2021).
29. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; 2020 Apr. 25–30; Honolulu. doi:10.1145/3313831.3376718.

Intérêts concurrents : Joseph Paul Cohen déclare avoir reçu une subvention Catalyseur en IA et une subvention Catalyseur IA-COVID-19 du CIFAR, ainsi qu'une subvention du Groupe de santé Carestream (Université Stanford), indépendamment des travaux soumis. Il déclare aussi être administrateur pour l'Institut de recherche sur les données reproductibles, un organisme sans but lucratif aux États-Unis. Michael Fralick déclare avoir reçu des honoraires de services-conseils de la société Proof Diagnostics (anciennement Pine Trees Health), une entreprise en démarrage qui travaille à la création d'un test diagnostique du virus SRAR-CoV2 utilisant la technologie CRISPR. Aucun autre intérêt concurrent n'a été déclaré.

Cet article a été révisé par des pairs.

Affiliations : Institut de recherche sur l'IA Mila (Cohen, Viviano, Huang, Bengio), Université de Montréal, Montréal, Qc.; Institut Vector sur l'intelligence artificielle (Cao, Ghassemi), Université de Toronto; Réseau hospitalier Unity Health de Toronto (Mamdani); Département de médecine (Fralick), Université de Toronto, Toronto, Ont.; Institut d'intelligence machine de l'Alberta (Greiner), Université de l'Alberta, Edmonton, Alb.; Département de radiologie (Cohen), et Centre d'intelligence artificielle machine & imagerie (Cohen), Université Stanford, Stanford, Calif.

Collaborateurs : Tous les auteurs ont contribué à la conception du travail, ont rédigé le manuscrit et en ont révisé de façon critique le contenu intellectuel important; ils ont donné leur approbation finale pour la version destinée à être publiée et assument l'entière responsabilité de tous les aspects du travail.

Correspondance : Joseph Paul Cohen, joseph@josephpcohen.com