Analyse

Évaluation des solutions fondées sur l'apprentissage machine en santé

Tony Antoniou PhD, Muhammad Mamdani PharmD MSP

■ Citation : *CMAJ* 2021. doi : 10.1503/cmaj.210036-f; diffusion hâtive le 30 août 2021

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

Les articles connexes présentent certains problèmes liés au développement de solutions fondées sur l'apprentissage machine en santé et suggèrent un cadre pour optimiser leur création[1,2]. L'utilisation de telles solutions est évaluée depuis quelques années dans de plus en plus de milieux cliniques. De nombreuses études présentent les données et les bases statistiques qui sont les fondements des outils fondés sur l'apprentissage machine[2], mais peu se sont intéressées à leur évaluation et à leur mise en œuvre[3]. Nous traiterons de l'évaluation des solutions fondées sur l'apprentissage machine tout au long de leur cycle de vie afin d'optimiser leur utilisation et leur utilité dans la pratique clinique. La validation interne — la vérification de la capacité de discrimination et de la correspondance avec un étalon des prédictions d'un algorithme — doit être suivie par l'évaluation du rendement et des résultats d'intérêts dans le milieu clinique ainsi que par l'évaluation de l'intégration de l'outil au déroulement du travail (comme indiqué à la figure 1).

Points clés

- L'évaluation des systèmes fondés sur l'apprentissage machine est un processus multifacettes qui comprend l'établissement de la validation interne et de la validation clinique, l'évaluation des résultats cliniques, la recherche sur la mise en œuvre et l'évaluation post-mise en œuvre.
- Les approches d'établissement de la validation clinique incluent la comparaison du rendement du modèle à celui d'experts de la santé et le déploiement silencieux de systèmes, et la comparaison des prédictions aux issues réelles pour les patients; l'évaluation des résultats cliniques peut se faire par des essais randomisés contrôlés, des études de cohortes, des études de séries temporelles interrompues et des études avant/après.
- La recherche sur la mise en œuvre comporte des éléments qualitatifs et quantitatifs et des évaluations formatives, et tient compte du contexte dans lequel le système sera déployé; les cadres d'évaluation peuvent aider les équipes à structurer leurs études et analyses.
- L'évaluation post-mise en œuvre est nécessaire pour surveiller l'apparition de menaces au rendement du système après son déploiement, menaces qui pourraient nécessiter une formation d'appoint ou une recalibration des systèmes.
- Une équipe multidisciplinaire composée de scientifiques des données, d'experts de la santé et de scientifiques de la mise en œuvre (expertise avec les données quantitatives et qualitatives) peuvent aider à assurer une évaluation complète avant, pendant et après le déploiement.

Quels sont les processus de création d'un modèle ou d'un algorithme et d'établissement de sa validation interne?

Pour évaluer le rendement prédictif des algorithmes fondés sur l'apprentissage machine, il faut d'abord évaluer leur capacité de discrimination et la correspondance de leurs prédictions avec un étalon. Le premier élément évalué quantifie la capacité de l'algorithme à classer les éléments selon la présence ou l'absence d'une caractéristique donnée; le second évalue dans quelle mesure les probabilités prédites par l'algorithme correspondent aux probabilités réelles[4]. Ces tests permettent d'évaluer la validation interne de l'algorithme et sont le sujet de la majorité des rapports publiés sur l'apprentissage machine en médecine[3].

Les études déterminant le rendement prédictif et l'exactitude de différents algorithmes sont généralement de nature rétrospective. De grands ensembles de données déjà étiquetées sont utilisés pour former et tester les algorithmes[3,5]. Les méthodes d'apprentissage machine utilisées à cette étape vont d'approches assez connues, comme la régression linéaire ou logistique, à des réseaux neuronaux plus complexes, en passant par les modèles de traitement du langage naturel[5,6]. Dans tous les cas, les algorithmes sont d'abord « formés » sur la plus grande portion des données réservées à cette fin, les données de formation, puis évalués à l'aide des données restantes, les données de test[3–5]. Lorsque le résultat d'intérêt est binaire (présence ou absence d'une maladie), le rendement est habituellement rapporté à l'aide de mesures typiques, comme la sensibilité, la spécificité et l'aire sous la courbe caractéristique du rendement[5,7]. Pour les résultats continus (prédiction de la dose d'un médicament), il est généralement quantifié à l'aide de mesures comme la racine carrée de l'erreur quadratique moyenne ou l'erreur quadratique absolue[8]. Les méthodes graphiques, comme les courbes d'étalonnage, peuvent être utilisées pour évaluer la calibration du modèle[9].

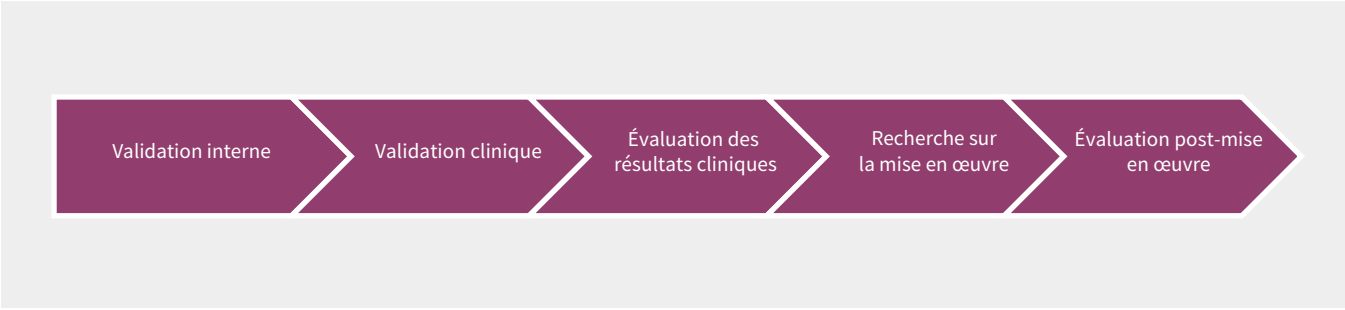

Figure 1 : Cycle d'évaluation des systèmes fondés sur l'apprentissage machine en santé.

À ce stade de développement technique, les commentaires de professionnels de la santé ou de parties prenantes ne semblent peut-être pas immédiatement nécessaires, mais ils peuvent fournir d'importantes informations sur l'interprétabilité des indicateurs de rendement et sur les seuils acceptables de rendement du modèle dans la pratique[10]. Par exemple, dans le cadre du développement d'un système d'alerte précoce fondé sur l'apprentissage machine prédisant la détérioration de l'état du patient et la nécessité de soins intensifs dans les 24 heures, le seuil d'acceptabilité a été établi par les professionnels de la santé à au plus 2 fausses alertes par véritable alerte afin de prévenir la « fatigue liée aux alertes »[1]. À partir de cette exigence, il a été déterminé que le système devrait avoir une valeur prédictive positive d'au moins 0,3 tout en détectant autant de patients dont l'état s'est détérioré que possible[1]. Parce que les valeurs optimales des indicateurs de rendement varient selon le contexte clinique, il faudra tenir compte des préférences des professionnels de la santé et du milieu de soins dans lequel le système sera utilisé pour les définir correctement.

Comment établir la validation clinique des solutions fondées sur l'apprentissage machine?

Le rendement des outils fondés sur l'apprentissage machine lorsqu'on leur présente de nouvelles données réelles pourrait différer du rendement pendant l'établissement de la validation interne[2]. Les études prospectives qui comparent les prédictions faites par l'algorithme à celles de professionnels de la santé doivent donc absolument vérifier son rendement en contexte clinique. Comme décrit dans un article connexe, cette approche a été utilisée dans le cadre de l'évaluation d'un système d'alerte précoce fondé sur l'apprentissage machine permettant de cibler les patients hospitalisés qui pourraient avoir besoin de soins intensifs; l'évaluation a révélé que le système d'alerte précoce avait une meilleure sensibilité que les professionnels de la santé[1]. Mentionnons aussi les exemples des comparaisons entre les professionnels de la santé et les systèmes fondés sur l'apprentissage machine pour le diagnostic de cancers de la peau[11–14]; du diagnostic de dégénérescence maculaire associée à l'âge et de rétinopathie diabétique à l'aide de tomographies par cohérence optique de la rétine ou de photographies du fond de l'œil[15–17]; du dépistage de métastases liées au cancer du sein dans les biopsies des ganglions lymphatiques[18,19]; et de la détection de polypes lors d'une coloscopie[20,21].

Une autre approche d'établissement de la validation clinique est la comparaison du rendement d'un algorithme nouvellement développé à celui des outils cliniques validés de calcul du risque couramment utilisés dans la pratique; cette approche a été appliquée à différents problèmes (prédiction de saignements gastro-intestinaux et de la mortalité après une chirurgie cardiaque[22,23]). Comme pour les approches exigeant une prédiction d'un professionnel de la santé, la comparaison avec les outils validés de calcul du risque doit se faire à l'aide de données ne faisant pas partie du processus de formation de l'algorithme.

Bien que de nombreuses études aient montré que le rendement des outils fondés sur l'apprentissage machine était au minimum comparable à celle de médecins experts, ce n'est pas toujours le cas[24], ce qui souligne le besoin de mener des études de vérification clinique avant d'entreprendre des formes d'évaluation plus coûteuses en ressources. L'établissement de la validation clinique peut être particulièrement difficile lorsque la fidélité interévaluateurs pour les professionnels de la santé émettant un diagnostic est faible. Dans ce contexte, il pourrait être ardu de comparer le rendement des professionnels de la santé à celui des systèmes fondés sur l'apprentissage machine en raison des défis associés à la discrimination entre la présence et l'absence d'une maladie ou les stades de la maladie (rémission, rechute). Des stratégies potentielles pour remédier à ce problème comprennent l'utilisation d'aspects concrets et mesurables d'une maladie donnée (changement dans la notation des symptômes ou dans les paramètres de laboratoire) ou un résultat fonctionnel directement observable (capacité à retourner au travail) plutôt que des étiquettes diagnostiques indiquant la présence ou l'absence de la maladie dans les données de formation.

Le « déploiement silencieux » est une autre approche qui peut être utilisée pour établir la validation clinique. Comme décrit dans un article connexe, dans cette approche, le système fondé sur l'apprentissage machine est exécuté et génère des prédictions, mais celles-ci ne sont pas communiquées au professionnel de la santé et n'influencent donc pas les soins[1]. Bien que le déploiement silencieux vise généralement à régler des problèmes liés au déploiement technique et au déroulement du travail, sans toucher aux interventions cliniques, les prédictions de l'outil pendant cette phase peuvent être comparées aux issues réelles pour les patients, ce qui permettrait d'estimer le rendement de l'algorithme.

Il ne faut habituellement pas de grands ensembles de données pour valider à l'avance les algorithmes d'apprentissage machine. La taille de l'échantillon peut être estimée à l'aide des méthodes établies pour l'évaluation de l'exactitude d'un test[25].

Comment déterminer si les solutions fondées sur l'apprentissage machine améliorent les issues pour les patients?

La vérification du rendement par des études sur la validation interne et clinique ne répond pas à une question fondamentale : l'intégration des solutions fondées sur l'apprentissage machine à la médecine clinique comporte-t-elle des avantages pour les patients[26]? Il est nécessaire de générer des données solides appuyant les retombées de ces algorithmes sur les issues pour les patients avant leur intégration étendue dans la pratique, et d'investir dans les ressources et les infrastructures nécessaires pour surveiller en continu le rendement de ces outils.

Comme c'est le cas pour d'autres types d'interventions, les essais randomisés contrôlés (ERC) sont la référence absolue pour déterminer l'efficacité des solutions fondées sur l'apprentissage machine. Pourtant, peu d'ERC sur ces solutions ont été enregistrées ou publiées[3,27]. On trouve un ERC à double insu sur un algorithme visant à détecter les complications neurologiques aiguës et un essai comparant l'effet de l'interprétation automatique des cardiotocographies à celle des soins traditionnels sur les issues cliniques chez les mères et les nourrissons[28,29]. La rareté des ERC en apprentissage machine peut s'expliquer par le besoin de grands échantillons de patients ou de longues durées de suivi pour montrer l'efficacité, les coûts et les problèmes relevant de la fidélité de l'intervention ou de la contamination entre les groupes lorsque les essais sont menés dans le même établissement. L'échantillonnage par grappes pourrait remédier à ce dernier problème, mais cette méthode ajoute à la complexité logistique et méthodologique déjà associée aux études multisites[30,31].

Parce qu'il est difficile de réaliser des ERC, d'autres approches sont souvent utilisées pour générer des données sur les avantages cliniques des systèmes fondés sur l'apprentissage machine, comme des études de cohorte appariée, des études de séries temporelles interrompues quasi expérimentales et des études prospectives avant/après[32–34]. Dans un article connexe, nous avons décrit comment nous prévoyons utiliser une étude de cohorte appariée observationnelle pour évaluer un système d'alerte précoce fondé sur l'apprentissage machine dans une unité de médecine interne générale, considérant qu'un ERC aurait nécessité un échantillon d'environ 30 000 patients[1]. Bien que les conclusions des études observationnelles soient souvent considérées de moins bonne qualité que les conclusions des ERC, elles représentent un compromis entre les besoins des parties prenantes et des professionnels de la santé d'obtenir des données récentes sur les retombées cliniques des interventions fondées sur l'apprentissage machine et les ressources nécessaires pour réaliser un ERC.

Comment optimiser la mise en œuvre des solutions fondées sur l'apprentissage machine?

Malgré le potentiel des interventions fondées sur l'apprentissage machine d'aider à la prise de décision clinique et d'améliorer le déroulement du travail, il n'existe présentement que quelques exemples de déploiement réussi en médecine[35]. De plus, peu d'études décrivent les étapes suivies pour transformer les algorithmes en outils cliniques. Pourtant, ces études sont cruciales pour cibler et éliminer les obstacles sociaux, éthiques, organisationnels et logistiques à l'adoption des solutions. La science de la mise en œuvre — l'étude des méthodes favorisant l'adoption d'une intervention dans la pratique — devrait donc être considérée comme aussi importante que la science des données et l'évaluation des issues cliniques pour l'intégration des systèmes fondés sur l'apprentissage machine en médecine[36,37]. Le présent article n'a pas pour objectif de décrire en détail la science de la mise en œuvre, mais plusieurs points méritent une attention particulière.

Contrairement aux études sur la validation interne et à la recherche clinique, qui se concentrent sur le rendement et l'efficacité des solutions, les questions de recherche sur la mise en œuvre et les issues se concentrent sur le processus de mise en œuvre, et peuvent inclure des mesures de l'adoption ou de l'acceptabilité d'une intervention; elles pourraient décrire la perception des fournisseurs par rapport à l'intégration de l'intervention au déroulement du travail ainsi que les changements à apporter aux processus de soins[37]. De plus, il est important de comprendre le contexte dans lequel le système est mis en œuvre pour en optimiser l'adoption[36]. Cela nécessite de se pencher sur des questions comme : comment intégrer le système au déroulement actuel du travail? Comment personnaliser l'interface utilisateur de manière à réduire au minimum les perturbations aux pratiques existantes? Quels membres de l'équipe de soins utiliseront le système?

La recherche sur la mise en œuvre peut utiliser une approche quantitative ou une approche qualitative. Les données quantitatives peuvent provenir de sondages structurés, de bases de données administratives en santé, des dossiers médicaux électroniques ou des systèmes d'aide à la prise de décision, selon les résultats à l'étude[38]. Des sondages peuvent être utilisés pour déterminer les éléments favorables et les obstacles à la mise en œuvre, pour connaître les attitudes quant à l'intégration d'un système dans le déroulement du travail établi et l'acceptabilité de l'intervention. Les dossiers médicaux peuvent être une source d'information sur l'adoption d'une intervention, la qualité des soins et les coûts. Les méthodes qualitatives peuvent quant à elles ajouter de la profondeur et du contexte aux approches quantitatives en étudiant comment et pourquoi une intervention est utilisée ou non par les professionnels de la santé, ce qui donne un aperçu des dynamiques interprofessionnelles et organisationnelles qui influencent l'adoption, ainsi que des obstacles socioculturels à la mise en œuvre[39]. Les données qualitatives peuvent être générées par des entrevues approfondies, des groupes de discussion, des analyses documentaires ou des observations, selon les questions de recherche et l'orientation méthodologique ou théorique des chercheurs.

Enfin, des évaluations formatives, dans lesquelles des données sont générées et remises à l'équipe de recherche et à certains professionnels de la santé à différentes étapes de la mise en œuvre, permettent à l'équipe de mise en œuvre de remédier aux problèmes qui surviennent et d'adapter la solution pour mieux l'intégrer au processus de soins[40]. L'utilisation d'un cadre d'évaluation ou d'une théorie pendant l'étude de la mise en

œuvre des outils fondés sur l'apprentissage machine peut aider les chercheurs à structurer leurs analyses et à préciser les concepts qui doivent être mesurés. Les lecteurs peuvent consulter d'autres sources expliquant les cadres d'évaluation couramment utilisés en science de la mise en œuvre[41].

Pourquoi l'évaluation post-mise en œuvre continue est-elle nécessaire?

Puisque la médecine et les processus cliniques évoluent avec le temps, l'évaluation des solutions fondées sur l'apprentissage machine ne se termine pas avec leur mise en œuvre. Il faut plutôt les évaluer continuellement pour en surveiller le rendement. Une menace importante au bon rendement de ces solutions est la désynchronisation de l'ensemble de données et des données réelles, qui survient lorsque des changements à la médecine clinique ou à la distribution des caractéristiques des patients mènent à un ensemble de données réelles qui diffère de l'ensemble de données de formation[42–44]. Cela peut se produire, par exemple, si un algorithme est utilisé pour faire des prédictions cliniques sur des données d'une population de plus en plus diversifiée sur le plan ethnique, ou dans un nouveau site avec une population de patients différant de l'ensemble de données de formation[2,45]. D'autres menaces au rendement des systèmes liées aux données pourraient être des changements par rapport aux variables utilisées lors de la phase de formation, comme l'ajout d'une nouvelle catégorie ou l'augmentation de la fréquence à laquelle certaines variables sont absentes.

L'évaluation continue du rendement du système peut nécessiter plusieurs étapes[46–49], dont une formation d'appoint du système avec les ensembles de données les plus récents, la comparaison du rendement du modèle lors de l'analyse des nouvelles données avec son rendement lors de l'analyse des données utilisées à ce moment et l'investigation des divergences; la mise à jour des définitions des résultats et des intrants du modèle pour suivre l'évolution de l'épidémiologie, du traitement ou des processus pathophysiologiques d'une maladie; la génération d'alertes déclenchées lorsque la fréquence de certaines variables change; et la consultation régulière d'experts pour détecter tout changement dans le rendement du système et en assurer la pertinence clinique. Lorsque possible, l'évaluation post-mise en œuvre d'une solution fondée sur l'apprentissage machine devrait être automatisée et prévue à intervalles réguliers pour détecter les sources de détérioration du système, enquêter sur ces problèmes et y remédier rapidement.

Conclusion

L'évaluation des solutions fondées sur l'apprentissage machine est un processus multifacettes qui nécessite l'expertise de scientifiques des données, d'experts de la santé et de scientifiques de la mise en œuvre. Présentement, la documentation décrivant l'évaluation de ces solutions demeure essentiellement axée sur la validation interne — assez peu d'études s'intéressent aux résultats cliniques et à la mise en œuvre des systèmes. Ce déséquilibre a contribué au fossé entre le développement et l'établissement de la validation des algorithmes et leur utilisation en médecine clinique[43]. D'autres études sur les résultats cliniques et la mise en œuvre sont donc nécessaires pour pleinement exploiter le potentiel de l'apprentissage machine en médecine.

Références

1. Verma AA, Murray J, Grenier R, et al. Implementing machine learning in medicine. *CMAJ* 2021;193:E1351-7.
2. Cohen JP, Cao T, Viviano JD, et al. Problems in the deployment of machine-learned models in health care. *CMAJ* 2021 Aug. 30 [cyberpublication avant impression]. doi:10.1503/cmaj.202066.
3. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.
4. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. 2nd ed. Leiden (Netherlands): Springer; 2019.
5. James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. New York: Springer; 2013.
6. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2:230-43.
7. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003;229:3-8.
8. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? — arguments against avoiding RMSE in the literature. *Geosci Model Dev* 2014;7: 1247-50.
9. Van Calster B, McLernon DJ, van Smeden M, et al. Topic group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
10. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. *JAMA* 2019;322:1351-2.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-8.
12. Han SS, Kim MS, Lim W, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol* 2018;138:1529-38.
13. Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. *J Am Acad Dermatol* 2020;82:622-7.
14. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018;29: 1836-42.
15. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122-31.e9.
16. Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol* 2017;135:1170-6.
17. Burlina P, Pacheco KD, Joshi N, et al. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. *Comput Biol Med* 2017;82:80-6.
18. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol* 2018;42:1636-46.
19. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-210.
20. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Ann Intern Med* 2018;169:357-66.
21. Chen PJ, Lin MC, Lai MJ, et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;154:568-75.
22. Allyn J, Allou N, Augustin P, et al. A comparison of a machine learning model with euroscore ii in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One* 2017;12:e0169772.

23. Shung DL, Au B, Taylor RA, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology* 2020;158:160-7.
24. Lehman CD, Wellman RD, Buist DS, et al.; Breast Cancer Surveillance Consortium. Diagnostic Accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-37.
25. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;7:371-92.
26. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med* 2018;1:40.
27. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
28. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med* 2018;24:1337-41.
29. Brocklehurst P, Field D, Greene K, et al. Computerised interpretation of the fetal heart rate during labour: a randomised controlled trial (INFANT). *Health Technol Assess* 2018;22:1-186.
30. Campbell MJ. Challenges of cluster randomized trials. *J Comp Eff Res* 2014;3:271-81.
31. Garrison MM, Mangione-Smith R. Cluster randomized trials for health care quality improvement research. *Acad Pediatr* 2013;13(Suppl 6):S31-7.
32. Bouaud J, Séroussi B, Antoine EC, et al. A before-after study using OncoDoc, a guideline-based decision support-system on breast cancer management: impact upon physician prescribing behaviour. *Stud Health Technol Inform* 2001;84:420-4.
33. Buising KL, Thursky KA, Black JF, et al. Improving antibiotic prescribing for adults with community acquired pneumonia: Does a computerised decision support system achieve more than academic detailing alone? — A time series analysis. *BMC Med Inform Decis Mak* 2008;8:35.
34. Harada Y, Shimizu T. Impact of a Commercial Artificial Intelligence-Driven Patient Self-Assessment Solution on Waiting Times at General Internal Medicine Outpatient Departments: Retrospective Study. *JMIR Med Inform* 2020;8:e21056.
35. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30-6.
36. Bauer MS, Damschroder L, Hagedorn H, et al. An introduction to implementation science for the non-specialist. *BMC Psychol* 2015;3:32.
37. Bauer MS, Kirchner J. Implementation science: What is it and why should I care? *Psychiatry Res* 2020;283:112376.
38. Smith JD, Hasan M. Quantitative approaches for the evaluation of implementation research studies. *Psychiatry Res* 2020;283:112521.
39. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. *Psychiatry Res* 2019;280:112516.
40. Elwy AR, Wasan AD, Gillman AG, et al. Using formative evaluation methods to improve clinical implementation efforts: description and an example. *Psychiatry Res* 2020;283:112532.
41. Nilsen P. Making sense of implementation theories, models and frameworks. *Implement Sci* 2015;10:53.
42. Beaulieu-Jones B, Finlayson SG, Chivers C, et al. Trends and focus of machine learning applications for health research. *JAMA Netw Open* 2019;2:e1914051.
43. Davis SE, Lasko TA, Chen G, et al. Calibration drift among regression and machine learning models for hospital mortality. *AMIA Annu Symp Proc* 2018;2017:625-34.
44. Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardiothorac Surg* 2013;43:1146-52.
45. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489-e92.
46. Toll DB, Janssen KJ, Vergouwe Y, et al. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085-94.
47. Siregar S, Nieboer D, Vergouwe Y, et al. Improved prediction by dynamic modeling: an exploratory study in the adult cardiac surgery database of the Netherlands association for cardio-thoracic surgery. *Circ Cardiovasc Qual Outcomes* 2016;9:171-81.
48. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691-8.
49. Janssen KJ, Moons KG, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76-86.

Intérêts concurrents : Aucun déclaré.

Cet article a été révisé par des pairs.

Affiliations : Centre de recherche et de formation en analytique des soins de santé Li Ka Shing (Antoniou, Mamdani), Réseau hospitalier Unity Health de Toronto; Institut du savoir Li Ka Shing (Antoniou, Mamdani), Réseau hospitalier Unity Health de Toronto; Département de médecine de famille et communautaire (Antoniou), Réseau hospitalier Unity Health deToronto et Université de Toronto; Faculté de médecine Temerty (Mamdani) et Faculté de pharmacie Leslie Dan (Mamdani), Université de Toronto; Institut des politiques, de la gestion et de l'évaluation de la santé (Mamdani), Université de Toronto, Toronto, Ont.

Collaborateurs : Les deux auteurs ont contribué à l'élaboration et à la conception des travaux. Tony Antoniou a rédigé l'ébauche du manuscrit. Muhammad Mamdani a révisé de façon critique le contenu intellectuel important du manuscrit. Les deux auteurs ont donné leur approbation finale pour la version soumise pour publication et assument l'entière responsabilité de tous les aspects du travail.

Correspondance : Tony Antoniou, tony.antoniou@unityhealth.to