

Mise en œuvre de l'apprentissage machine en santé

Amol A. Verma MD MAPH, Joshua Murray MSc, Russell Greiner PhD, Joseph Paul Cohen PhD, Kaveh G. Shojania MD MSc, Marzyeh Ghassemi PhD, Sharon E. Straus MD MSc, Chloé Pou-Prom MSc, Muhammad Mamdani PharmD MSP

■ Citation : *CMAJ* 2021 August 30;193:E1351-7. doi : 10.1503/cmaj.202434-f

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

L'apprentissage machine — le développement de systèmes qui, à partir de données, apprennent à reconnaître des tendances et à faire des prédictions justes d'événements à venir¹ — a un fort potentiel pour transformer le domaine de la santé. Les outils fondés sur l'apprentissage machine pourraient appuyer la prise de décision clinique complexe et permettre d'automatiser de nombreuses tâches usuelles qui peuvent faire perdre du temps aux professionnels de la santé et provoquer une insatisfaction par rapport au travail². Malgré l'intérêt grandissant envers de telles technologies, comme les algorithmes des montres intelligentes détectant la fibrillation auriculaire³, et leur approbation réglementaire qui prend de l'ampleur, leur utilisation en médecine clinique est demeurée jusqu'à maintenant limitée⁴. Le développement et la mise en œuvre de ces outils en santé nécessitent une infrastructure et des ressources qui peuvent être difficiles d'accès — des ensembles de données cliniques réelles de grande taille, des compétences techniques en sciences des données, et une capacité et une infrastructure informatiques majeures. Les enjeux entourant la sécurité et la confidentialité des données, le mauvais rendement des modèles mathématiques, la difficulté à intégrer les outils au déroulement du travail, la faible acceptabilité des solutions fondées sur l'apprentissage machine chez les professionnels de la santé et l'incertitude entourant la façon d'évaluer ces solutions peuvent être des obstacles à leur adoption⁴. Nous proposons ici une approche au développement et à l'adoption des solutions fondées sur l'apprentissage machine en santé. Les articles connexes traitent de certaines restrictions à l'utilisation de ces technologies⁵ et de leur évaluation⁶.

Le développement de solutions fondées sur l'apprentissage machine à des fins cliniques nécessite une solide compréhension des soins cliniques, de la science des données et de la science de la mise en œuvre. Il existe d'excellents cadres appuyant les initiatives d'analytique des données et d'amélioration de la qualité, notamment le processus pansectoriel normalisé de forage de données (Cross-Industry Standard Process for Data Mining — CRISP-DM)⁷, le modèle d'amélioration de l'Institut pour l'amélioration de soins de santé⁸ et le cadre d'application des connaissances (Knowledge to

Points clés

- L'apprentissage machine a le potentiel de transformer le domaine de la santé, mais ses applications en médecine clinique sont pour le moment limitées.
- Les partenariats multidisciplinaires entre experts techniques et utilisateurs (professionnels de la santé, gestionnaires, patients et membres de la famille) sont essentiels au développement et à la mise en œuvre des solutions fondées sur l'apprentissage machine en santé.
- Un cadre en 3 phases peut être utilisé pour décrire le développement et l'adoption des solutions fondées sur l'apprentissage machine : une phase d'exploration pour comprendre le problème à régler et l'environnement de déploiement, une phase de conception de la solution pour le développement de modèles fondés sur l'apprentissage machine et d'outils conviviaux, et une phase de mise en œuvre et d'évaluation pour le déploiement de la solution et l'évaluation de ses retombées.

Action Framework)⁹. Cela dit, aucun cadre clair et complet ne vise l'adoption des outils fondés sur l'apprentissage machine en santé. Nous proposons donc un cadre en 3 phases (exploration, conception de la solution, et mise en œuvre et évaluation; figure 1), étayé par un exemple concret (encadré 1), pour favoriser l'adoption de ces solutions en médecine clinique. Le cadre peut être utilisé avec différentes solutions, comme des projets de vision artificielle, d'automatisation et d'optimisation, et d'analytique prédictive. Il peut aussi être appliqué lors de la mise en œuvre d'une solution développée par un tiers, puisque les étapes, à l'exception du développement du modèle, restent les mêmes.

Principales étapes de la phase d'exploration

Le développement de solutions fondées sur l'apprentissage machine fonctionnelles nécessite une compréhension approfondie du problème à régler, des résultats pertinents, des données disponibles au moment du développement et dans l'avenir,

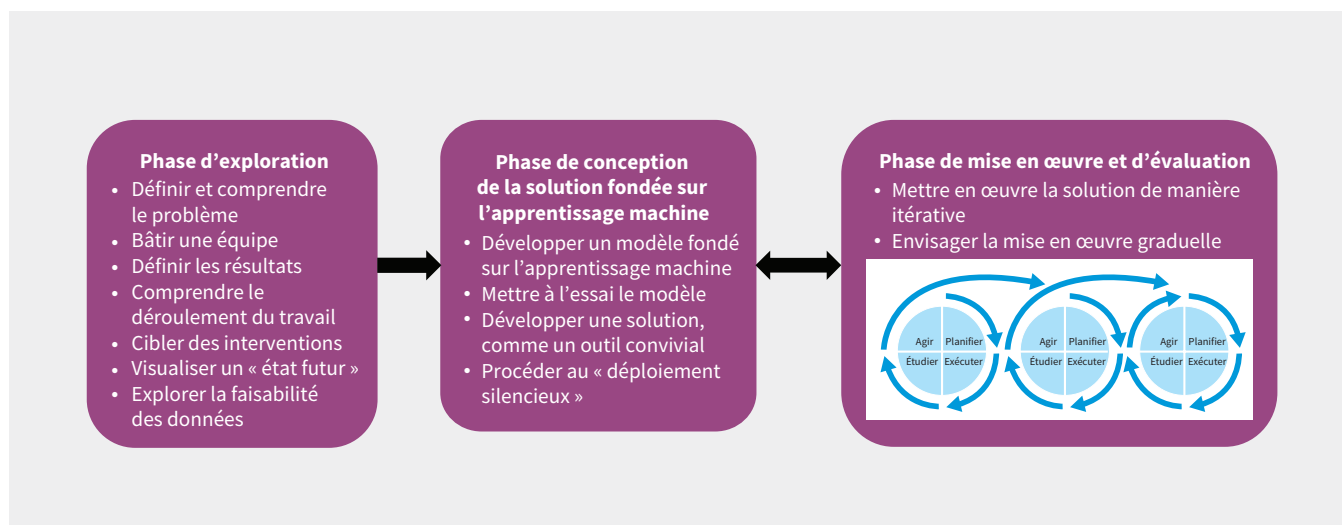


Figure 1 : Cadre de développement et d'adoption des solutions fondées sur l'apprentissage machine en médecine clinique.

des besoins des utilisateurs, du déroulement du travail, des facteurs humains et de la gestion du changement. La mise en œuvre des solutions d'aide à la prise de décision clinique peut être renforcée par une compréhension, dès le début, de l'association entre la solution et une intervention clinique éprouvée pour améliorer les soins.

Encadré 1 : Exemple concret

L'incapacité de reconnaître la détérioration de l'état clinique dans les hôpitaux est l'une des principales causes de transfert imprévu des patients aux unités de soins intensifs¹⁰. Les systèmes d'alerte précoce^{11,12} peuvent prédire le risque de détérioration de l'état clinique d'un patient et potentiellement permettre aux professionnels de la santé d'intervenir plus tôt. De nombreux systèmes d'alerte précoce existants se fondent sur des approches statistiques traditionnelles, comme les modèles de régression logistique utilisant une combinaison simple d'un petit nombre d'intrants (le plus souvent, moins de 10 paramètres, comme les signes vitaux), et ils sont enclins aux faux-positifs¹³. Les modèles biostatistiques plus avancés pourraient cibler de manière plus exacte les patients à risque¹³. Malgré tout, la mise en œuvre et l'évaluation des modèles biostatistiques ou des modèles fondés sur l'apprentissage machine est peu courante.

Le service de médecine interne générale de l'Hôpital St. Michael, un centre universitaire de santé à Toronto (Ontario), traite environ 4000 patients chaque année. Approximativement 7 % des patients du service décèdent ou sont transférés à l'unité des soins intensifs¹⁴. L'hôpital a une équipe de soins intensifs bien rodée (composée d'un inhalothérapeute ainsi que d'une infirmière et d'un médecin spécialisés en soins intensifs) qui peut être appelée par les services pour des soins urgents aux patients hospitalisés qui pourraient nécessiter un transfert à l'unité de soins intensifs. L'hôpital a commencé en 2017 à développer un système d'alerte précoce fondé sur l'apprentissage machine pour le service de médecine interne générale. L'objectif était de prédire et de prévenir la détérioration de l'état clinique des patients pour réduire le taux de mortalité. La mise en œuvre et l'évaluation de l'intervention, qui s'est déroulée de manière itérative en 2020, sont en cours.

Délimiter le problème et bâtir une équipe

La première étape est de trouver un problème important aux yeux des utilisateurs, comme les professionnels de la santé ou les gestionnaires, et de définir les résultats mesurables précis qu'ils souhaitent changer en modifiant les pratiques actuelles. Les solutions fondées sur l'apprentissage machine pourraient viser le remplacement d'une tâche manuelle; dans ce cas, le résultat serait le gain de temps et les mesures de rendement pour la tâche en question. Les solutions pourraient aussi être conçues pour remédier à un problème clinique; le résultat serait alors une amélioration clinique mesurable. Les problèmes sont habituellement d'abord repérés par les utilisateurs, et devraient ensuite être étudiés par une équipe multidisciplinaire, qui déterminera si une solution fondée sur l'apprentissage machine pourrait être appropriée. L'équipe devrait être composée d'utilisateurs qui comprennent le problème clinique ou opérationnel et le déroulement du travail, d'ingénieurs des données et de professionnels des technologies de l'information (TI) qui connaissent les données et l'infrastructure disponibles et qui savent comment y intégrer une solution, de scientifiques des données qui savent développer des modèles fondés sur l'apprentissage machine, et de patients et de proches aidants lorsque les solutions proposées les visent.

Comme le développement et la mise en œuvre de solutions fondées sur l'apprentissage machine nécessitent beaucoup de ressources, il faut choisir avec attention les projets à prioriser. D'abord, le problème doit être majeur, ce qui pourrait être déterminé en estimant à quel point sa résolution améliorera la santé des patients, leur expérience ou celle des fournisseurs de soins, ou à quel point elle réduira les coûts. Ensuite, une solution d'apprentissage machine doit être réalisable, ce qui est déterminé par la disponibilité de données de qualité en quantité suffisante dans un délai raisonnable, par la probabilité de réussite d'un modèle pour ce problème, et par la capacité de mettre en œuvre une solution qui s'intégrera bien dans l'infrastructure informatique existante et le déroulement du travail clinique. Enfin, les interventions qui accompagneront la solution doivent

avoir une probabilité raisonnable d'amélioration. Idéalement, les interventions proposées sont fondées sur des données probantes et leur efficacité a déjà été confirmée. Au bout du compte, la participation des utilisateurs est essentielle au succès du projet. Les utilisateurs n'adopteront une solution fondée sur l'apprentissage machine que si elle s'intègre au déroulement de leur travail et qu'elle se révèle utile.

Comprendre le problème et établir des objectifs

Les utilisateurs peuvent signaler un problème qu'ils rencontrent régulièrement, mais ils pourraient ne pas comprendre pourquoi le problème existe ou comment le résoudre. L'équipe multidisciplinaire est là pour analyser le problème et élaborer une théorie de changement qui décrit sa meilleure hypothèse quant à l'amélioration qui pourrait être apportée par une solution fondée sur l'apprentissage machine. Des approches systématiques à la compréhension des problèmes cliniques et opérationnels ont été bien décrites, notamment par la schématisation des processus, l'analyse des relations de cause à effet, et l'analyse des modes de défaillance et de leurs effets¹⁵. La compréhension du problème

oriente le développement, la mise en œuvre et l'évaluation de la solution. Comme pour tout projet d'amélioration, l'équipe devrait établir des objectifs d'amélioration clairs et mesurables en définissant les résultats pertinents, en décrivant le rendement au point de référence et en établissant un objectif d'amélioration précis. L'équipe devrait aussi établir un indice de référence de rendement du modèle, concept propre aux solutions fondées sur l'apprentissage machine, qui serait utile dans la pratique clinique. Il pourrait être constructif de répondre à la question : quel est le rendement actuel des décideurs, et à quel point doit-il être amélioré pour qu'une solution fondée sur l'apprentissage machine en vaille la peine? Un modèle ayant une forte exactitude qui ne donne pas de meilleurs résultats que le jugement clinique sera moins utile qu'un modèle moyennement exact qui donne de bien meilleurs résultats que le jugement clinique.

Dans l'exemple présenté à l'encadré 1, une équipe d'exploration (figure 2) a été assemblée pour envisager les différents événements cliniques qui pourraient être prédits (sepsis, insuffisance rénale aiguë, réadmission, etc.) pour améliorer les soins aux patients dans un service de médecine interne générale.

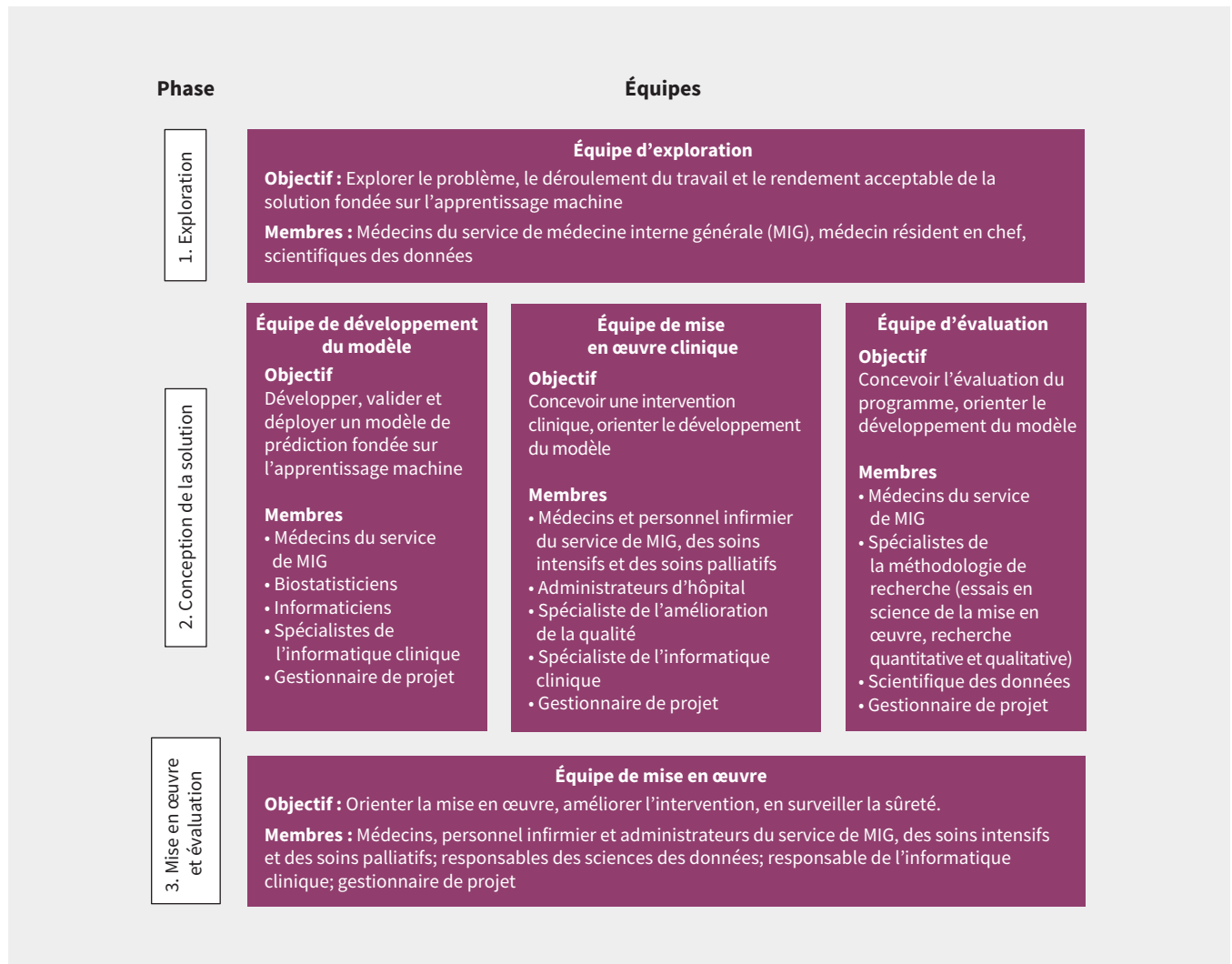


Figure 2 : Structure des équipes pour chaque phase de développement du système d'alerte précoce pour le service de médecine interne générale (MIG) de l'Hôpital St. Michael, Toronto (Ontario). Remarque : USI = unité de soins intensifs.

Selon l'analyse documentaire et les données disponibles, l'équipe a créé une liste d'options retenues, puis consulté tous les membres du service, l'administrateur de l'hôpital et 3 des conseillers aux patients et aux membres de la famille de l'hôpital avant de choisir la détérioration de l'état clinique (décès ou transfert à l'unité des soins intensifs) comme priorité. Les experts des données et des TI ont déterminé que le projet était faisable. L'analyse documentaire, les discussions avec les médecins et le personnel infirmier du service et un court examen des dossiers de 10 cas de détérioration de l'état clinique choisis au hasard¹⁶ ont permis à l'équipe de mieux comprendre le problème. La théorie de changement proposée était qu'un système d'alerte fondé sur l'apprentissage machine pourrait améliorer les soins en permettant la détection précoce de la détérioration de l'état clinique, ce qui permettrait aux professionnels de la santé d'intervenir plus tôt, d'avoir des discussions proactives sur les préférences et les objectifs de soins du patient et d'accélérer la consultation des équipes de soins intensifs ou de soins palliatifs. L'équipe s'est donné pour objectif de réduire la mortalité chez les patients admis au service de médecine interne générale de 10 % en 1 an, ce qui était considéré comme atteignable compte tenu d'autres études sur les systèmes d'alerte précoce¹⁷.

Conception des solutions fondées sur l'apprentissage machine

Le développement d'une solution fondée sur l'apprentissage machine demande la conception et la mise à l'essai d'un modèle fondé sur l'apprentissage machine puis sa mise en œuvre préliminaire. Nous suggérons l'utilisation d'un cadre comme le CRISP-DM⁷ pour le développement et la mise à l'essai des algorithmes. Un grand avantage de cette approche est qu'elle tient compte de la nature itérative de la science des données, qui nécessite souvent 6 phases : l'analyse du cas, l'analyse des données, la préparation des données, la création du modèle, l'évaluation de rendement du modèle, et le déploiement du modèle. Cette approche intègre plusieurs facteurs à considérer (problème auquel il faut remédier; quantité, qualité et type de données disponibles; facteurs à considérer dans la mise en œuvre — déroulement du travail et acceptabilité chez les utilisateurs). Le développement d'une solution fondée sur l'apprentissage machine nécessite souvent 3 flux de travail complémentaires, qui pourraient être dirigés par 1 ou plusieurs équipes : développement du modèle, mise en œuvre dans le milieu et évaluation du modèle (figure 2). Ces flux de travail sont interdépendants, puisque les décisions prises dans un flux ont des répercussions sur les 2 autres. Des équipes ciblées peuvent être mises sur pied pour chaque flux de travail afin que chacun reçoive une attention et une expertise suffisantes, et des membres devraient faire partie de plus d'une équipe pour assurer une bonne coordination.

Vérifier la qualité des données

De nombreux problèmes rencontrés lors du déploiement d'une solution fondée sur l'apprentissage machine sont dus aux données utilisées pour le développement du modèle. La qualité des intrants peut être évaluée par des approches automatisées rela-

tivement simples et des validations manuelles ciblées¹⁹ pour en déterminer l'exhaustivité, l'exactitude, la concordance, la plausibilité et l'actualité¹⁸. Outre ces indicateurs de qualité de base, il est aussi important de comprendre les résultats utilisés pour former le modèle et de déterminer s'ils correspondent réellement aux prédictions désirées. Un des articles connexes traite des problèmes liés aux données de formation⁵.

Concevoir le modèle en pensant à sa mise en œuvre

Il existe différents moyens de développer des modèles efficaces, notamment les techniques de régression traditionnelles, comme les régressions logistiques, et les techniques d'apprentissage machine plus modernes qui peuvent traiter les relations complexes entre les variables, comme les réseaux neuronaux¹. Les scientifiques des données choisissent une approche adaptée aux caractéristiques des intrants et des résultats²⁰, mais toute la solution devrait être conçue par une équipe multidisciplinaire qui a en tête sa mise en œuvre²¹. Dans l'exemple (encadré 1), la solution impliquait un modèle de prédiction, un système de communication du niveau de risque du patient aux professionnels de la santé et un parcours de soins pour les patients chez qui le risque était élevé. Tous les éléments de la solution ont été conçus de manière itérative par les 3 équipes (figure 2), avec des commentaires périodiques des conseillers aux patients et aux membres de la famille. Les équipes ont décidé qu'afin de maintenir un équilibre entre le temps nécessaire pour évaluer les patients chez qui le risque est élevé et le temps nécessaire aux autres tâches, le modèle de prédiction devrait émettre au plus 2 fausses alertes pour chaque véritable alerte. Les scientifiques des données ont donc établi le seuil de classification des patients à risque élevé à une valeur prédictive positive de 30 %, selon les données existantes. À ce seuil, la sensibilité était de 50 %, ce que les professionnels de la santé considéraient comme une proportion utile de cas détectés. Ces professionnels ont jugé qu'il serait le plus utile de prédire les issues qui se produiraient probablement dans les 24 à 48 heures. Une fenêtre plus courte ne leur laisserait pas le temps d'intervenir, et une fenêtre plus longue rendrait le choix de l'intervention difficile. Les scientifiques des données ont donc entraîné le modèle à prédire les complications dans les 48 prochaines heures.

Développer un outil convivial

Pour les systèmes d'aide à la prise de décision, les modèles devraient être intégrés au sein du déroulement du travail, dans des outils conviviaux qui donnent des informations clés utiles et utilisables. Cela nécessite la collaboration entre les utilisateurs et les experts en amélioration des procédés, en facteurs humains, en conception et en gestion du changement. La participation des utilisateurs est essentielle tout au long du processus, mais son ampleur dépend du problème à régler. Dans l'exemple, selon les principes des facteurs humains²², une approche simple en 3 catégories a été choisie pour présenter de l'information utile aux professionnels de la santé — les patients sont classés selon leur niveau de risque : élevé, modéré ou faible. Les prédictions sont envoyées par l'outil électronique de

consignation des informations à communiquer lors du caucus de changement de quart et par des notifications sur le téléavertisseur des professionnels de la santé. Les notifications ne sont transmises que lorsque le risque des patients passe de faible ou modéré à élevé; si le risque d'un patient demeure élevé, aucune autre notification n'est envoyée, ce qui réduit la fatigue liée aux alertes²³. Il y a ainsi entre 0 et 2 alertes par période de 24 heures par équipe de soins (une équipe s'occupe habituellement de 15–20 patients).

Concevoir une intervention clinique à intégrer au déroulement du travail

L'introduction d'un nouvel outil clinique, qu'il soit fondé sur l'apprentissage machine ou non, peut modifier le déroulement du travail²⁴. Ces changements peuvent être planifiés et bien accueillis²⁵, ou dérangeants et nuisibles²⁶. Différentes stratégies, notamment des entrevues, des groupes de discussion, des sondages et des analyses du déroulement du travail, permettent de décrire le déroulement du travail de référence et de mettre en évidence les éléments favorables et les obstacles à l'adoption d'un nouvel outil^{24,27}. Les informations ainsi obtenues peuvent ensuite être schématisées dans des stratégies efficaces pour optimiser la mise en œuvre à l'aide

d'approches comme le modèle COM-B (capacité, opportunité, motivation, comportement)²⁸. Dans l'exemple, l'équipe de mise en œuvre était composée de professionnels de la santé et de gestionnaires ayant une expérience concrète du déroulement du travail au service de médecine interne générale, de soins intensifs et de soins palliatifs ainsi que du système informatique utilisé. Au besoin, on a tenu des entrevues et des groupes de discussion supplémentaires pour orienter le travail de l'équipe de mise en œuvre. Pour concevoir l'intervention, l'équipe a tenu compte des ressources existantes, comme les protocoles hospitaliers pour l'intensification des soins et ceux de l'équipe de soins intensifs. Le mode et le moment de transmission des alertes ont été pensés pour s'intégrer aux procédures existantes pour les médecins et le personnel infirmier des services de médecine interne générale, de soins intensifs et de soins palliatifs. En effet, les prédictions du modèle sont transmises au personnel infirmier responsable à des moments précis et dans un format défini pour que le niveau de risque des patients soit pris en compte dans l'assignation du travail. Un parcours clinique a été conçu, qui propose des actions concrètes et des cibles de délai de réponse pour les patients à risque élevé tout en laissant la place au jugement clinique (figure 3).

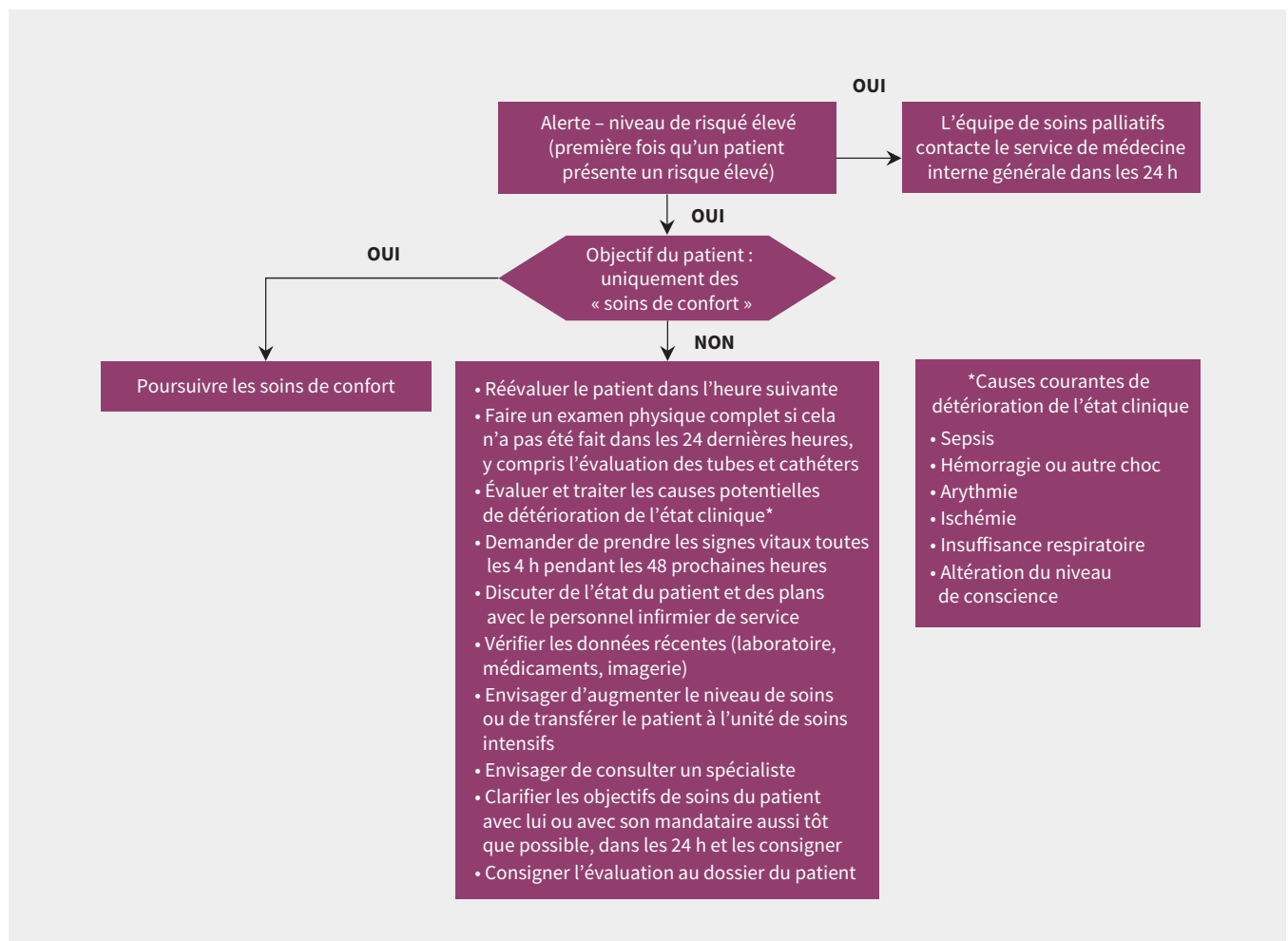


Figure 3 : Parcours des patients dont le risque de détérioration de l'état clinique prédit est élevé dans le service de médecine interne générale.

Faire participer les utilisateurs pour établir un lien de confiance

Un obstacle courant à l'adoption des technologies fondées sur l'apprentissage machine est le manque de confiance des professionnels de la santé envers les résultats²⁹. Un cadre suggère que la démonstration de la transparence, de l'exactitude et de la solidité des modèles peut inspirer confiance³⁰. Dans l'exemple, l'équipe a utilisé des données réelles obtenues entre 2011 et 2020 pour développer et valider le modèle du système d'alerte précoce. Des modèles de régression multivariée par spline adaptative ont été développés à l'aide d'environ 100 intrants sur les caractéristiques démographiques, les signes vitaux et les résultats de test de laboratoire des patients; ce type de modèle a été choisi après la mise à l'essai de différentes techniques de modélisation à l'aide de plus de 500 variables³¹. L'important nombre d'intrants et leurs interactions complexes rendent difficile l'explication des facteurs qui influencent les prédictions, mais certains modèles sont plus faciles à interpréter que d'autres (il est possible de rapporter l'importance relative de différents facteurs de prédiction). L'interprétabilité des modèles peut être souhaitable dans certaines applications cliniques³², mais elle n'est pas nécessaire pour inspirer la confiance³³, et il n'y a pas de consensus sur les meilleures méthodes pour expliquer les modèles plus complexes^{34,35}. Les explications détaillées sur les prédictions d'un modèle pourraient même nuire à la prise de décision dans certains cas, par une surcharge d'information ou en créant une fausse impression de causalité.

Pour inspirer confiance envers le système d'alerte précoce du service de médecine interne générale, nous avons expliqué de manière transparente aux professionnels de la santé de première ligne comment nous avons développé et validé les modèles fondés sur l'apprentissage machine, en leur montrant qu'ils ne comportent aucun biais lié à l'âge ou au sexe du patient (peu de données sociodémographiques permettaient toutefois d'explorer d'autres dimensions de l'exactitude). Nous avons démontré la solidité du système en effectuant une validation des modèles par partition d'échantillon sur une base temporelle à l'aide de données réelles : les modèles formés sur les données de 2011 à 2019 ont été mis à l'essai avec les données de 2020. Nous avons aussi comparé les prédictions du système à celles des médecins et du personnel infirmier pendant 4 mois pour établir sa validité clinique et son utilité potentielle. Pour faciliter la participation des utilisateurs, l'initiative a été chapeautée par des dirigeants cliniques reconnus, notamment les responsables des soins infirmiers et les médecins responsables des services de médecine interne générale, de soins intensifs et de soins palliatifs.

Il est important de faire participer les patients, les membres de la famille et les proches aidants, surtout pendant le développement de solutions qui serviront directement les patients. On peut ainsi améliorer la conception et la sûreté des nouveaux services ainsi que la satisfaction par rapport à ceux-ci^{36,37}. Les méthodes de ce type de participation ont été amplement décrites^{38,39} et devraient comprendre une définition claire de l'objectif de la mobilisation, des adaptations aux besoins uniques pour rendre accessible la participation, du processus de recrutement de partenaires diversifiés, et de l'acceptation des

occasions d'échange entre les personnes ayant des connaissances théoriques et celles ayant une expérience concrète. Dans l'exemple, le recrutement des patients et des proches aidants s'est fait principalement dans les groupes consultatifs de l'hôpital, et on les a consultés à différentes étapes du projet. Nous avons choisi un modèle de participation consultative afin de solliciter des commentaires sur les enjeux clés, notamment la sélection de la détérioration de l'état clinique comme une priorité, la conception de l'intervention clinique et la recherche de solutions aux problèmes liés à la mise en œuvre. Par exemple, un grand sujet de discussion était la façon d'informer les patients et leurs proches des prédictions du modèle. L'équipe de mise en œuvre clinique a conclu des discussions que les médecins devraient être responsables de discuter des prédictions du modèle avec le patient lorsque cliniquement approprié et d'intégrer ces prédictions au plan de santé et de traitement.

Mise en œuvre et évaluation des solutions fondées sur l'apprentissage machine

Mise en œuvre graduelle

Il n'est pas recommandé d'adopter largement les solutions fondées sur l'apprentissage machine en santé immédiatement après leur développement. Il devrait plutôt y avoir une période de « déploiement silencieux » avant la mise en œuvre officielle, période pendant laquelle les utilisateurs ne sont pas informés des prédictions du modèle. La durée de cette période est déterminée par plusieurs facteurs, dont la fréquence des prédictions, la nature de la pratique ciblée, et le nombre et la diversité des utilisateurs prévus. Cette période permet de vérifier que les données et l'infrastructure informatique fonctionnent bien et que le rendement du modèle dans le monde réel est suffisant pour la mise en œuvre. Si le déploiement silencieux est fructueux, les résultats peuvent être présentés aux utilisateurs pour renforcer la confiance. Dans le cas contraire, la phase de mise à l'essai peut prévenir le déploiement d'un modèle potentiellement nuisible ou mettre en lumière les améliorations nécessaires. Dans l'exemple, le modèle a été déployé en mode silencieux pendant 9 mois. Nous avons repéré plusieurs problèmes, que nous avons réglés; par exemple, nous avons corrigé une erreur de traitement où l'algorithme interprétait « Na » (symbole chimique du sodium) comme « NA » (valeur manquante), ce qui nuisait au rendement du modèle.

Évaluation itérative

En raison de la complexité du milieu de la santé et du processus de développement d'un modèle, nous suggérons l'adoption d'une approche itérative utilisant le cycle « planifier – exécuter – étudier – agir »^{40,41} décrit dans le modèle d'amélioration de l'Institut pour l'amélioration de soins de santé⁸. Pour ce faire, il faut « planifier » le déploiement de la solution, ses objectifs et les mesures clés de son efficacité et de sa sécurité; « exécuter » la mise en œuvre à petite échelle; « étudier » le processus de mise en œuvre et ses retombées sur les mesures établies; et « agir » pour améliorer le processus de mise en œuvre selon l'étape « étudier ». L'évaluation de la mise en œuvre des modèles fondés sur l'apprentissage machine est un processus itératif — décrit

plus en profondeur dans un article connexe⁶ — qui nécessite souvent plusieurs cycles avant que la solution ne soit intégrée efficacement au déroulement du travail.

Dans l'exemple, après le déploiement silencieux, nous avons lancé le système d'alerte précoce lors d'un déploiement graduel dans 2 équipes cliniques du service de médecine interne générale en août 2020, puis nous l'avons étendu aux 5 équipes du service en septembre, et enfin au personnel infirmier et à l'équipe de soins palliatifs en octobre. L'approche graduelle nous a permis de repérer et de corriger tout problème imprévu en lien avec le modèle, l'environnement informatique ou le déroulement du travail. Pendant la mise en œuvre, les 3 équipes responsables des phases d'exploration et de conception de la solution ont été fusionnées en 1 seule équipe de mise en œuvre (figure 2) qui se rencontrait toutes les semaines pour vérifier les mesures du processus et des résultats afin d'améliorer l'intervention et l'adhésion au parcours clinique et pour remédier aux conséquences imprévues. Nous avons corrigé les erreurs, comme les messages d'alerte erronés, révisé les critères des alertes et modifié le processus de formation pour les médecins et le personnel infirmier.

Méthodes d'évaluation

Bien qu'un essai randomisé contrôlé (ERC) soit l'idéal pour étudier les retombées des interventions, il est possible d'opter pour d'autres types d'études, comme les études de séries temporelles interrompues. Dans l'exemple, la réalisation d'un ERC a été envisagée, mais la taille de l'échantillon nécessaire a rayé cette option de la liste — il aurait fallu environ 30 000 participants pour détecter une réduction relative du taux de mortalité de 10%, le taux de mortalité réel étant de 6%. Une approche pragmatique à méthode mixte a plutôt été adoptée, ce qui comprend une évaluation qualitative pour cibler les obstacles à la mise en œuvre et étudier les effets de la solution sur la pratique par des entrevues approfondies avec le personnel infirmier, les résidents et les médecins. Les méthodologies d'une étude de séries temporelles interrompues et d'une étude par cohorte appariée seront utilisées pour comparer les issues pour les patients dans la période d'intervention aux données du groupe témoin, recueillies par le passé. Ces 2 approches pourraient limiter les facteurs de confusion à long terme et à l'échelle du patient, mais les facteurs de confusion de la pandémie de COVID-19 demeureront une limite importante. Des réseaux d'essais multisites voués à l'évaluation des nouvelles technologies fondées sur l'apprentissage machine sont nécessaires pour permettre une évaluation rigoureuse.

Conclusion

L'idée selon laquelle l'apprentissage machine pourrait rapidement et dramatiquement transformer le milieu de la santé par l'automatisation des tâches usuelles et l'amélioration de la prise de décision clinique est séduisante. Malheureusement, la réalité de l'apprentissage machine en santé est toute autre — il y a eu de nombreux cas d'échec de la mise en œuvre de tels outils⁵. Pour trouver des solutions fondées sur l'apprentissage machine qui fonctionnent, il faut travailler prudemment avec le « désordre » des données de santé et la complexité des décisions cliniques et du déroulement du travail. L'apprentissage machine a un potentiel

considérable de faire avancer fortement le milieu de la santé, mais une approche disciplinée, inclusive, déterminée et itérative au développement et à l'adoption de ces technologies est nécessaire pour que les patients en bénéficient réellement.

Références

1. Liu Y, Chen PHC, Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806-16.
2. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016;165:753-60.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56.
4. Ben-Israel D, Jacobs WB, Casha S, et al. The impact of machine learning on patient care: a systematic review. *Artif Intell Med* 2020;103:101785.
5. Cohen JP, Cho 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.
6. Antoniou T, Mamdani MM. Evaluation of machine learning solutions in medicine. *CMAJ* 2021 Aug. 30 [cyberpublication avant impression]. doi:10.1503/cmaj.210036.
7. Chapman P, Clinton J, Kerber R, et al. CRISP-DM 1.0: a step-by-step data mining guide. Armonk (NY): SPSS; 2000. Accessible ici : <https://www.the-modeling-agency.com/crisp-dm.pdf> (consulté le 18 mai 2021).
8. How to improve. Boston: Institute for Healthcare Improvement. Accessible ici : <http://www.ihp.org/resources/Pages/HowtoImprove/default.aspx> (consulté le 18 mai 2021).
9. Graham ID, Logan J, Harrison MB, et al. Lost in knowledge translation: time for a map? *J Contin Educ Health Prof* 2006;26:13-24.
10. van Galen LS, Struik PW, Driesen BEJM, et al. Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: A root cause analysis of unplanned ICU admissions. *PLoS One* 2016;11:e0161393. doi: 10.1371/journal.pone.0161393.
11. Burch VC, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and in-hospital mortality. *Emerg Med J* 2008;25:674-8.
12. McGinley A, Pearse RM. A national early warning score for acutely ill patients. *BMJ* 2012;345:e5310.
13. Linnen DT, Escobar GJ, Hu X, et al. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. *J Hosp Med* 2019;14:161-9.
14. Verma AA, Guo Y, Kwan JL, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open* 2017;5:E842-9.
15. Quality Improvement Essentials Toolkit. Boston: Institute for Healthcare Improvement; 2021. Accessible ici : www.ihp.org/resources/Pages/Tools/Quality-Improvement-Essentials-Toolkit.aspx (consulté le 18 mai 2021).
16. Etchells E, Ho M, Shojania KG. Value of small sample sizes in rapid-cycle quality improvement projects. *BMJ Qual Saf* 2016;25:202-6.
17. Escobar GJ, Liu VX, Schuler A, et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020;383:1951-60.
18. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144-51.
19. Verma AA, Pasricha SV, Jung HY, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *J Am Med Inform Assoc* 2021;28:578-87.
20. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317-8.
21. Shah NH, Milstein A, Bagley Steven CP. Making machine learning models clinically useful. *JAMA* 2019;322:1351-2.
22. Phansalkar S, Edworthy J, Hellier E, et al. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *J Am Med Inform Assoc* 2010;17:493-501.
23. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. *AACN Adv Crit Care* 2013;24:378-86.
24. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record-supported work to improve health system performance. *Ann Intern Med* 2020;172(Suppl 11):S116-22.

25. Teich JM, Merchia PR, Schmitz JL, et al. Effects of computerized physician order entry on prescribing practices. *Arch Intern Med* 2000;160:2741-7.
26. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;293:1197-203.
27. Craig P, Dieppe P, Macintyre S, et al. *Developing and evaluating complex interventions: the new Medical Research Council guidance*. *BMJ* 2008;337:a1655.
28. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci* 2011;6:42.
29. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;46:50-80.
30. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22:e15154.
31. Nestor B, McCoy LG, Verma AA, et al. Preparing a clinical support model for silent mode in general internal medicine. Proceedings of the 5th Machine Learning for Healthcare Conference, *PMLR* 2020;126:950-72. Accessible ici : https://static1.squarespace.com/static/59d5ac1780bd5ef9c396eda6/t/5f22ccec4a74012a31d4b4a/1596116209243/155_CameraReadySubmission_155_nestor2020preparing.pdf (consulté le 14 sept. 2021).
32. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2:749-60.
33. Lipton ZC. The myths of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 2018;16:1-28. doi: 10.1145/3236386.3241340.
34. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc* 2020;27:1173-85.
35. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics* 2019;8:832.
36. Bombard Y, Baker GR, Orlando E, et al. Engaging patients to improve quality of care: a systematic review. *Implement Sci* 2018;13:98. doi: 10.1186/s13012-018-0784-z.
37. Sharma AE, Knox M, Mleczko VL, et al. The impact of patient advisors on healthcare outcomes: a systematic review. *BMC Health Serv Res* 2017;17:693. doi: 10.1186/s12913-017-2630-4.
38. Kiran T, Tepper J, Gavin F. Working with patients to improve care. *CMAJ* 2020;192: E125-7.
39. Hamilton C, Hoens AM, Backman CL, et al. Workbook to guide the development of a patient engagement in research (PEIR) plan. Richmond (BC): Arthritis Research Canada and Vancouver: University of British Columbia; 2018. Accessible ici : <http://www.arthritisresearch.ca/wp-content/uploads/2018/06/PEIR-Plan-Guide.pdf> (consulté le 9 juin 2021).
40. Taylor MJ, McNicholas C, Nicolay C, et al. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290-8.
41. Leis JA, Shojania KG. A primer on PDSA: executing plan-do-study-act cycles in practice, not just in name. *BMJ Qual Saf* 2017;26:572-7.
42. Blier N. Stories of AI failure and how to avoid similar AI fails [blog]. Amherst (MA): Lexalytics; 2020 Jan 30. Accessible ici : <https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020> (consulté le 18 mai 2021).

Intérêts concurrents : Amol Verma déclare avoir reçu une bourse de recherche en compassion et intelligence artificielle de l'organisme AMS Healthcare, dans le cadre des travaux présentés ici. Le Dr Verma signale aussi avoir reçu une subvention de projet exploratoire de l'Institut Vecteur (pour les travaux présentés ici) et est employé à temps partiel par l'organisme Santé Ontario (qui n'a joué aucun rôle dans les travaux présentés ici). Amol Verma, Joshua Murray, Chloé Pou-Prom et Muhammad Mamdani ont dirigé le développement et la mise en œuvre du système d'alerte précoce CHARTwatch à l'Hôpital St. Michael, Toronto, Ontario, Canada. Aucun autre intérêt concurrent n'a été déclaré.

Cet article a été révisé par des pairs.

Affiliations : Réseau hospitalier Unity Health de Toronto (Verma, Murray, Straus, Pou-Prom, Mamdani); Institut du savoir Li Ka Shing de l'Hôpital St. Michael (Verma, Straus, Pou-Prom, Mamdani); Département de médecine (Verma, Shojania, Straus, Mamdani) et Institut des politiques, de la gestion et de l'évaluation de la santé (Verma, Mamdani) et Département de statistique (Murray), Université de Toronto, Toronto, Ont.; Université de l'Alberta (Greiner); Institut d'intelligence machine de l'Alberta (Greiner), Edmonton, Alb.; Institut des algorithmes d'apprentissage de Montréal (Cohen), Montréal, Qc.; Centre pour l'amélioration de la qualité et la sécurité des patients (Shojania), Université de Toronto; Centre des sciences de la santé Sunnybrook (Shojania); Institut Vecteur (Ghassemi, Mamdani) et Département des sciences informatiques (Ghassemi); Faculté de pharmacie Leslie Dan (Mamdani), Université de Toronto, Toronto, Ont.; Département de radiologie, Université Stanford (Cohen), 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.

Propriété intellectuelle du contenu : Il s'agit d'un article en libre accès distribué conformément aux modalités de la licence Creative Commons Attributions (CC BY-NC-ND 4.0), qui permet l'utilisation, la diffusion et la reproduction dans tout médium à la condition que la publication originale soit adéquatement citée, que l'utilisation se fasse à des fins non commerciales (c.-à-d., recherche ou éducation) et qu'aucune modification ni adaptation n'y soit apportée. Voir : <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr>.

Correspondance : Muhammad Mamdani, muhammad.mamdani@unityhealth.to; et Amol Verma, amol.verma@mail.utoronto.ca