https://doi.org/10.24095/hpcdp.42.1.04f

Recherche quantitative originale

Une approche d'apprentissage automatique pour prédire l'utilisation des cigarettes électroniques et la dépendance à celles-ci chez les jeunes de l'Ontario

Jiamin Shi, M.S.P. (1,2); Rui Fu, M. Sc. (2,3); Hayley Hamilton, Ph. D. (1,2); Michael Chaiton, Ph. D. (1,2)

Cet article a fait l'objet d'une évaluation par les pairs.

Résumé

Introduction. Nous avons mis au point des algorithmes de forêt aléatoire pour prédire le risque que les jeunes Ontariens essaient un jour la cigarette électronique (vapotage) et qu'ils l'utilisent de façon quotidienne, puis nous avons examiné l'importance des prédicteurs et l'interaction statistique.

Méthodologie. Cette étude transversale repose sur un échantillon représentatif d'élèves du primaire et du secondaire de l'Ontario en 2019 (N = 6471). Nous avons utilisé la fréquence du vapotage au cours des 12 derniers mois pour définir l'essai de vapotage (avoir déjà expérimenté de vapoter) et le vapotage quotidien. Nous avons intégré un vaste ensemble de caractéristiques individuelles comme corrélats possibles de l'essai de vapotage (176 variables) et du vapotage quotidien (179 variables). À l'aide de la validation croisée, nous avons élaboré des algorithmes de forêt aléatoire et nous avons évalué le rendement du modèle selon l'indice de concordance, une mesure qui permet d'évaluer la capacité discriminatoire d'un modèle, et ce, pour les deux résultats. En outre, nous avons défini les 10 corrélats principaux grâce au calcul du score de l'importance relative et leur interaction avec les caractéristiques sociodémographiques.

Résultats. Dans l'échantillon, 2064 (31,9 %) répondants avaient déjà essayé le vapotage et 490 (7,6 %) des répondants étaient des consommateurs quotidiens. Les algorithmes de forêt aléatoire pour les deux résultats ont fourni une performance élevée, avec un indice de concordance supérieur à 0,90. Les 10 corrélats principaux du vapotage quotidien concernaient la consommation de caféine, de cannabis et de tabac, la source et le type de cigarette électronique et l'absentéisme scolaire au cours des 20 derniers jours. Les corrélats de l'essai de vapotage étaient la taille de l'école, la consommation d'alcool, de cannabis et de tabac et, de plus, 9 des 10 principaux corrélats de l'essai de vapotage affichaient des interactions avec l'ethnicité.

Conclusion. L'apprentissage automatique est une méthodologie prometteuse pour déterminer les risques d'essai de vapotage et de vapotage quotidien. En outre, il permet d'en cerner les corrélats importants et d'évaluer les recoupements complexes, ce qui pourrait être utile pour les futures études longitudinales visant à personnaliser les politiques de santé publique pour certains groupes cibles de population.

Mots clés : *apprentissage automatique, vapotage, tabagisme, Ontario, jeunes*

Points saillants

- Cette étude a appliqué une méthodologie d'apprentissage automatique qui a permis d'inclure un large éventail de corrélats dans la recherche sur le tabagisme chez les jeunes.
- Les 10 corrélats principaux du vapotage quotidien concernent la consommation de caféine, de cannabis et de tabac, la source et le type de cigarette électronique et l'absentéisme scolaire au cours des 20 derniers jours. Les corrélats de l'essai de vapotage étaient la taille de l'école et la consommation d'alcool, de cannabis et de tabac.
- Les futures études longitudinales pourraient vérifier les corrélats observés en matière d'essai de vapotage et de vapotage quotidien, ce qui serait sans doute utile pour l'adoption de stratégies ciblant en priorité les problèmes liés à la consommation de substances.
- L'analyse des interactions a permis de quantifier la force de ces interactions parmi les principaux corrélats et les caractéristiques sociodémographiques, ce qui pourrait être approfondi dans de futures études longitudinales.

Rattachement des auteurs :

1. École de santé publique Dalla Lana, Université de Toronto, Toronto (Ontario), Canada
2. Institut de recherche sur les politiques en santé mentale/Centre de toxicomanie et de santé mentale, Toronto (Ontario), Canada
3. Institut des politiques, de la gestion et de l'évaluation de la santé, École de santé publique Dalla Lana, Université de Toronto, Toronto (Ontario), Canada

Correspondance : Michael Chaiton, Institut de recherche sur la santé mentale des populations, Centre de toxicomanie et de santé mentale, Toronto (Ontario) M5T 1R8; tél. : 416-535-8501; courriel : michael.chaiton@camh.ca

Introduction

La recherche a montré une augmentation rapide de la prévalence du vapotage de produits à base de nicotine chez les jeunes nord-américains de 16 à 19 ans entre 2017 et 2018[1]. En particulier, le pourcentage de jeunes ayant déjà essayé le vapotage est passé de 29,3 % à 37,0 %, et le pourcentage de jeunes ayant vapoté au cours des 30 derniers jours est passé de 8,4 % à 14,6 % au Canada. Les jeunes déclarent également de plus en plus qu'ils ressentent des symptômes de dépendance au vapotage, définis comme « la multitude de comportements et de symptômes qui provoquent un état de détresse chez l'utilisateur et l'incitent à faire une utilisation compulsive du vapotage en raison de facteurs liés ou non à la nicotine » [TRADUCTION][2, p. 257]. Une étude de cohorte prospective laisse penser que la dépendance au vapotage pourrait être associée à la persistance future et à l'intensification future du tabagisme chez les élèves de 12e année aux États-Unis[3]. En 2020, environ 3 000 hospitalisations et décès signalés par les Centers for Disease Control and Prevention (CDC) des États-Unis étaient associés à l'utilisation de produits de vapotage[4].

Des études antérieures sur la dépendance au vapotage, dont certaines utilisant des échelles validées comme PROMIS-E et l'indice de dépendance à la cigarette électronique de Penn State, ont attribué l'augmentation des symptômes de dépendance au vapotage à un âge plus avancé, à une durée d'utilisation supérieure, à une fréquence de vapotage supérieure, à des concentrations de nicotine élevées et à l'usage de la cigarette[5,6]. Toutefois, ces études comportent des limites associées aux régressions statistiques classiques. L'utilisation de valeurs *p* pour sélectionner les caractéristiques du modèle proposé en fonction de la signification statistique peut limiter la compréhension des prédicteurs non sélectionnés. De plus, comme la dépendance au vapotage peut être corrélée à tout un éventail de caractéristiques, on peut difficilement saisir l'ensemble de ces relations complexes dans un modèle de régression. Cette complexité risque de limiter les résultats de l'étude en raison de problèmes statistiques comme la multicolinéarité et le surajustement.

Pour tenir compte de ces limites, nous avons appliqué pour cette étude une approche d'apprentissage automatique. L'apprentissage automatique – une technique émergente employée dans la recherche en santé[8-11] – correspond à « un groupe de méthodes analytiques axées sur les données qui reposent sur la puissance computationnelle pour exécuter des fonctions statistiques »[7, p. 1317]. Par rapport aux méthodes statistiques classiques, l'apprentissage automatique offrirait une meilleure exactitude prédictive, avec des lignes directrices appropriées pour atténuer les risques de surajustement[12]. La définition de « prédicteur » que nous utilisons ici est celle employée dans le domaine de l'apprentissage automatique, en référence à un modèle de prédiction, et elle n'implique pas de relation temporelle ou causale.

Notre méthodologie met l'accent sur les variables qui sont les plus « importantes » pour la prédiction en matière d'amélioration du rendement de l'aire sous la courbe (ASC) de la courbe de la fonction d'efficacité du récepteur (ROC) du modèle, de préférence aux estimations de la variance et à la valeur *p* d'un test d'hypothèse. Bien que certaines études aient appliqué des méthodes d'apprentissage automatique comme les arbres de classification[13] et les forêts aléatoires[14] dans la recherche sur le tabagisme, une récente étude de portée a révélé que ces applications étaient rarement en lien avec des résultats en santé publique[15]. Dans ce contexte, notre étude avait pour but d'analyser l'essai de vapotage et le vapotage quotidien (comme indicateur de dépendance au vapotage) au sein d'une population de jeunes à l'aide de méthodes d'apprentissage automatique donnant lieu à des observations interprétables. Plus particulièrement, nos objectifs étaient d'élaborer des algorithmes d'apprentissage automatique qui prédisent à la fois l'essai de vapotage et le vapotage quotidien chez les jeunes Ontariens et d'effectuer des analyses ultérieures, notamment en hiérarchisant l'importance des facteurs de risque individuels à l'égard des deux résultats et en illustrant les intersections statistiques pour identifier les sous-groupes de jeunes particulièrement vulnérables.

Méthodologie

Données et participants

Cette étude repose sur les données du Sondage sur la consommation de drogues et la santé des élèves de l'Ontario (SCDSEO) réalisé en 2019, qui a fourni les réponses de 14 142 élèves répartis dans 992 classes de 263 écoles primaires ou secondaires appartenant à 47 conseils scolaires de l'Ontario[16]. Le SCDSEO a un plan d'échantillonnage complexe, les écoles étant regroupées selon 26 strates géographiques. Le sondage comptait quatre types de questionnaires. Nous avons obtenu un ensemble de 6471 répondants après avoir inclus uniquement les questionnaires intégrant la question « Au cours des 12 derniers mois, à quelle fréquence avez-vous fumé des cigarettes électroniques? », en excluant les élèves n'ayant pas répondu à cette question. Seuls les répondants ayant déjà essayé le vapotage, soit 2 064 répondants, ont été intégrés à l'échantillon utilisé pour analyser le vapotage quotidien.

Mesures

Résultat

Nous avons créé des variables binaires correspondant au vapotage quotidien et à l'essai de vapotage en utilisant la même question de sondage. Les répondants ayant déclaré n'avoir jamais utilisé de ci-garette électronique ont été classés comme n'ayant jamais vapoté et les autres comme ayant déjà vapoté. Les participants vapotant au moins une fois par jour ont été considérés comme ayant une dépendance au vapotage. Ceux qui ne répondaient pas à ce critère ont été considérés comme des répondants ne vapotant pas quotidiennement.

Déterminants potentiels

Nous avons utilisé 179 variables relatives aux caractéristiques individuelles susceptibles de prédire un vapotage quotidien et 176 variables pour l'essai de vapotage[16] (voir l'annexe à https://osf.io/x36p8/ pour la liste complète des variables). Ces variables intègrent des données administratives, des données démographiques, la vie scolaire, la vie familiale, la santé physique, la santé mentale, les comportements au volant, le fait d'avoir été passager d'un véhicule conduit par une personne en état d'ébriété, les comportements liés au vapotage, la consommation de substances, les perceptions et l'exposition, les caractéristiques sociodémographiques et d'autres comportements à risque liés à la consommation de substances. Nous avons exclu toutes les variables conditionnelles au vapotage quotidien ou à l'essai de vapotage du fait de la structure du questionnaire (les questions impliquant d'avoir fait l'essai du vapotage n'ont pas été incluses comme

prédicteurs de l'essai du vapotage). Nous avons regroupé les niveaux de plusieurs variables afin de faciliter l'analyse subséquente. Nous avons mis à l'échelle les variables numériques en utilisant la normalisation par score z avant de construire le modèle.

Analyses statistiques

Statistiques descriptives et imputation des valeurs manquantes

Nous avons effectué une synthèse des caractéristiques individuelles des répondants et des prévalences de l'essai de vapotage et du vapotage quotidien. Plus de 90 % des variables avaient des données manquantes dans une proportion inférieure à 5 % ou comprise entre 5 % et 10 %. Une variable décrivant différents types d'éducation spécialisée avait 10 % de données manquantes. Les variables nominales ont soit été regroupées avec leurs niveaux de référence, soit présentées comme des options représentant l'incertitude quant à la façon de répondre. Nous avons imputé la valeur manquante à la médiane pour toutes les variables numériques.

Algorithme de forêt aléatoire

À l'aide du logiciel R version 3.6.3 « caret »[17], nous avons élaboré un algorithme de forêt aléatoire – un algorithme d'apprentissage automatique d'ensemble formé par un grand nombre d'arbres de classification – pour classer les répondants en fonction des résultats primaires[18]. Par exemple, dans l'algorithme du vapotage quotidien, chaque arbre a classé les répondants comme étant vapoteurs quotidiens ou comme n'étant pas vapoteurs quotidiens. Après l'addition de toutes les prédictions de classe des arbres, la classe ayant obtenu la majorité de votes est devenue la prédiction de la forêt aléatoire. Cette approche axée sur la « sagesse des foules » a le potentiel de faire de la forêt aléatoire un algorithme extrêmement précis et rigoureux pour la prédiction[19].

Élaboration et validation d'une forêt aléatoire pour le vapotage quotidien et l'essai de vapotage

Nous avons inclus tous les prédicteurs candidats pour entraîner le modèle, à l'exclusion des variables dépendant du résultat (exclusion des questions portant sur l'essai de vapotage posées aux élèves qui vapotaient). En utilisant un ratio de 7:3, nous avons réparti aléatoirement l'ensemble de données en un ensemble d'entraînement (n = 1612 ou 4680) et un ensemble d'essai (n = 691 ou 2006) pour l'échantillon afin de classer le vapotage quotidien et l'essai de vapotage. Les deux formes de vapotage étaient déséquilibrées. Pour faciliter l'efficacité de l'entraînement du modèle, nous avons effectué une procédure SMOTE (technique de suréchantillonnage des minorités synthétiques) sur les données d'entraînement afin d'atteindre deux échantillons équilibrés pour l'entraînement du modèle[20]. Dans le cadre d'une procédure de validation croisée à 10 blocs pendant l'entraînement du modèle, l'ensemble de données a été réparti aléatoirement en 10 sous-échantillons de taille égale. À chaque itération, neuf sous-échantillons ont été utilisés pour entraîner le modèle, tandis que le sous-échantillon retenu a été utilisé pour valider le modèle. La procédure ci-dessus a été répétée 10 fois. Pour évaluer le rendement du modèle, nous avons noté l'exactitude, la sensibilité, la spécificité et l'ASC relativement à la classification du vapotage quotidien et de l'essai de vapotage pour l'ensemble d'essai. Nous avons considéré que le rendement moyen des 10 itérations était le rendement global du modèle. L'ASC dépassant 0,80 a fourni une bonne capacité discriminatoire, un seuil habituel pour ce modèle de classification[21].

Classement des facteurs de risque individuels du vapotage quotidien et de l'essai de vapotage

Pour déterminer les 10 corrélats principaux du vapotage quotidien et de l'essai de vapotage, nous avons classé tous les corrélats en fonction des scores de l'importance relative mis à l'échelle (0 à 100), une mesure calculée à partir de la perte totale d'exactitude due à l'exclusion d'un corrélat pour chaque arbre divisée par le nombre total d'arbres[22,23]. Des graphiques de la dépendance partielle unidirectionnelle des 10 corrélats principaux ont été utilisés pour comprendre leurs effets marginaux sur les risques prévus de vapotage quotidien et d'essai de vapotage, les autres corrélats demeurant constants[24]. Le graphique de dépendance partielle d'un corrélat a illustré les probabilités des résultats en fonction des valeurs différentes de ce corrélat. Plus la probabilité était élevée, plus le risque de résultat observé sous l'influence de ce corrélat était élevé. Ces méthodes ont également été appliquées aux caractéristiques sociodémographiques.

Exploration des interactions

Nous avons examiné les interactions bidirectionnelles des 10 corrélats principaux identifiés et des corrélats sociodémographiques pouvant prédire avec robustesse les inégalités quant aux résultats liés au tabagisme[25]. De plus, nous avons exploré les effets d'interaction des paires de caractéristiques sociodémographiques suivantes – âge et sexe, âge et ethnicité, âge et statut socioéconomique (SSE), sexe et ethnicité, sexe et SSE, ethnicité et SSE – à l'aide d'un système de classement simple mesurant l'importance des caractéristiques[26]. Les répondants avaient déterminé subjectivement leur SSE sur une échelle allant de 0 à 10[27]. Des graphiques de dépendance partielle bidirectionnelle ont été utilisés pour illustrer les risques de vapotage quotidien et d'essai de vapotage dans les paires proposées, avec des forces d'interaction supérieures à un seuil de 0,1. Les calculs des probabilités de dépendance partielle ont été fondés sur la variation des deux prédicteurs alors que les autres prédicteurs demeuraient constants[28].

Analyse de sensibilité

Nous avons effectué deux ensembles d'analyses de sensibilité en utilisant le même ensemble d'entraînement suréchantillonné pour les deux résultats. Nous avons d'abord ajusté les algorithmes de forêt aléatoire avec les 10 corrélats principaux uniquement. Nous avons ensuite élaboré des modèles de régression logistique à variables multiples de base comprenant l'âge, le sexe, l'ethnicité et le SSE. Le rendement de ces modèles logistiques a été évalué en fonction de l'exactitude, de la sensibilité, de la spécificité et de l'ASC dans l'ensemble d'essai et il a été comparé aux mesures de la forêt aléatoire.

Résultats

Caractéristiques de l'échantillon

Les 6471 répondants ont été répartis en 10 groupes d'âge (0 à 11 ans, chaque année entre 12 et 19 ans et 20 ans et plus); 54,6 % d'entre eux étaient de sexe féminin; la majorité (68,6 %) était issue de familles ayant une cote de 6 à 8 sur l'échelle du SSE et 62,1 % d'entre eux étaient d'ethnicité blanche (tableau 1). Parmi les répondants, 2064 (31,9 %) avaient déjà vapoté au moins une fois et 490 (soit 7,6 % de l'échantillon complet et 23,7 % de ceux ayant déjà vapoté) vapotaient quotidiennement.

Rendement des algorithmes de forêt aléatoire

Les algorithmes de forêt aléatoire ont fourni une performance élevée pour les

TABLEAU 1
Caractéristiques individuelles des répondants du SCDSEO 2019 inclus dans l'échantillon

	Données globales (N = 6471)
Âge (ans)	
11 ou moins	20 (0,3 %)
12	727 (11,2 %)
13	954 (14,7 %)
14	1042 (16,1 %)
15	1225 (18,9 %)
16	1100 (17,0 %)
17	981 (15,2 %)
18	386 (6,0 %)
19	27 (0,4 %)
20 ou plus	9 (0,1 %)
Sexe	
Féminin	3535 (54,6 %)
Masculin	2936 (45,4 %)
Statut socioéconomique[a]	
1	6 (0,1 %)
2	40 (0,6 %)
3	122 (1,9 %)
4	280 (4,3 %)
5	675 (10,4 %)
6	1061 (16,4 %)
7	1805 (27,9 %)
8	1575 (24,3 %)
9	657 (10,2 %)
10	250 (3,9 %)
Ethnicité	
Blanche	4017 (62,1 %)
Chinoise	374 (5,8 %)
Sud-Asiatique	648 (10,0 %)
Noire	563 (8,7 %)
Autochtone	157 (2,4 %)
Philippine	368 (5,7 %)
Amérique latine/Amérique centrale/Amérique du Sud	282 (4,4 %)
Asiatique du Sud-Est	125 (1,9 %)
Asiatique proche-occidentale ou arabe	344 (5,3 %)
Coréenne	56 (0,9 %)
Japonaise	31 (0,5 %)
Incertitude sur l'origine ethnique	256 (4,0 %)
Expérience de vapotage	
Non	4407 (68,1 %)
Oui	2064 (31,9 %)
Vapotage quotidien	
Non	5981 (76,3 %)
Oui	490 (23,7 %)

Abréviations : SCDSEO, Sondage sur la consommation de drogues et la santé des élèves de l'Ontario, SSE, statut socioéconomique.

[a] Le SSE a été déterminé subjectivement par les répondants en fonction de la cote qu'ils accordaient à leur propre SSE sur l'échelle de MacArthur du statut socioéconomique subjectif, une échelle allant de 0 à 10.

deux résultats. L'algorithme pour l'essai de vapotage a eu une exactitude d'essai de 0,82 (intervalle de confiance [IC] à 95 % : 0,81 à 0,84), une sensibilité de 0,83 (0,80 à 0,86), une spécificité de 0,82 (0,80 à 0,84) et une ASC de 0,90. L'algorithme pour le vapotage quotidien a eu une exactitude d'essai de 0,83 (0,80 à 0,86), une sensibilité de 0,85 (0,77 à 0,90), une spécificité de 0,82 (0,78 à 0,86) et une ASC de 0,90.

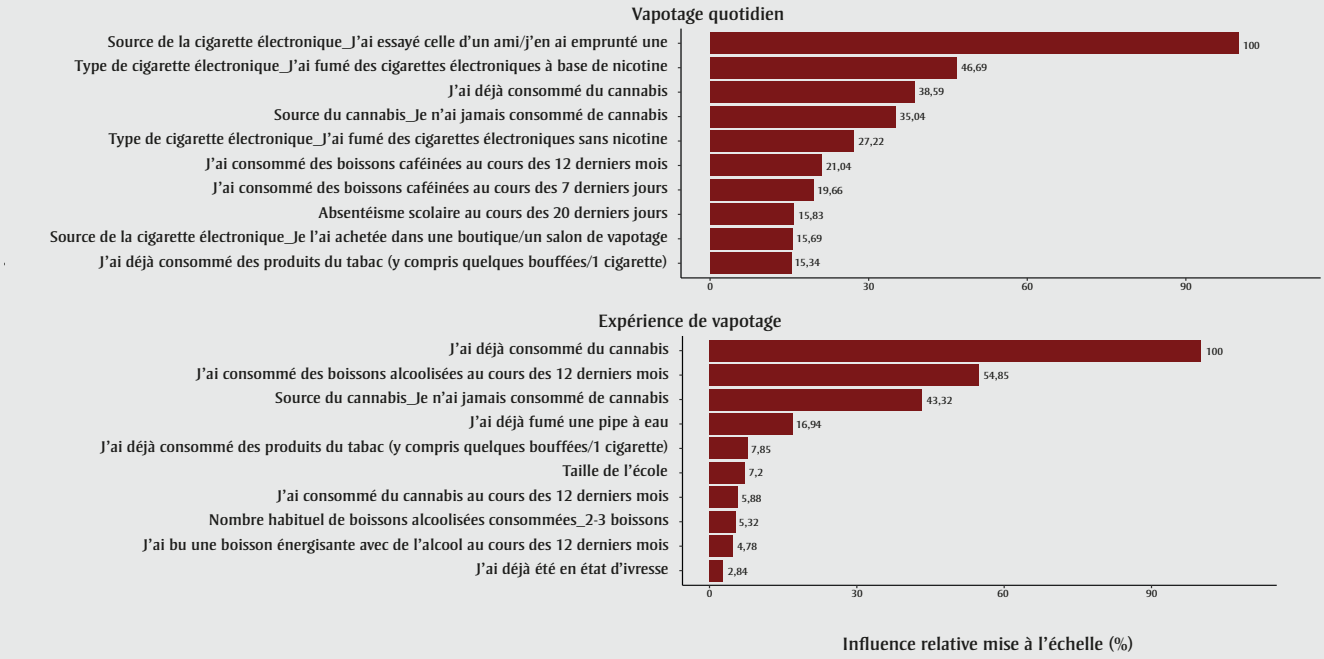
10 corrélats principaux de l'essai de vapotage et du vapotage quotidien

Les algorithmes ont fourni 10 corrélats principaux pour le vapotage quotidien et 10 pour l'essai de vapotage (figure 1). Pour l'essai de vapotage, ces 10 corrélats principaux étaient : avoir déjà consommé du cannabis; avoir consommé de l'alcool au cours des 12 derniers mois; la source de cannabis; avoir déjà essayé la pipe à eau; avoir déjà consommé du tabac; la taille de l'école; avoir consommé du cannabis au cours des 12 derniers mois; le nombre de boissons alcoolisées consommées habituellement; avoir consommé une boisson énergisante avec de l'alcool au cours des 12 derniers mois et enfin avoir été en état d'ivresse. Les 10 corrélats principaux du vapotage quotidien étaient comme source de cigarette électronique, avoir essayé celle d'un ami; avoir fumé des cigarettes électroniques à base de nicotine; avoir déjà consommé du cannabis; la source de cannabis; avoir fumé des ci-garettes électroniques sans nicotine; avoir consommé une boisson caféinée au cours des 12 derniers mois; avoir consommé une boisson caféinée au cours des 7 derniers jours; absentéisme scolaire au cours des 20 derniers jours; comme source de cigarette, avoir acheté des cigarettes électroniques dans une boutique de vapotage et enfin avoir déjà consommé des produits du tabac. Tant pour le vapotage quotidien que pour l'essai de vapotage, étant donné que tous les corrélats sociodémographiques présentaient une influence minimale avec une importance relative inférieure à trois, nous ne présentons pas les graphiques de dépendance partielle correspondants.

Dépendance partielle pour les 10 prédicteurs principaux

D'après les graphiques de dépendance partielle se rapportant à l'essai de vapotage, nous avons constaté un risque supérieur, par rapport aux répondants ne l'ayant pas fait, chez les répondants ayant

FIGURE 1
Importance relative remise à l'échelle des 10 corrélats principaux du vapotage quotidien et de l'essai de vapotage, SCDSEO 2019



Abréviation : SCDSEO, Sondage sur la consommation de drogues et la santé des élèves de l'Ontario.

consommé du cannabis au cours des 12 derniers mois ou au cours de leur vie, ayant bu de l'alcool avec ou sans boisson énergisante au cours des 12 derniers mois, ayant consommé du tabac ou fumé une pipe à eau au cours de leur vie et ayant déjà été en état d'ivresse (voir l'annexe à https://osf.io/x36p8/). En matière de source de cannabis, les répondants ayant déjà consommé du cannabis présentaient un risque plus élevé d'avoir fait l'essai du vapotage que ceux n'en ayant jamais consommé. Les répondants consommant habituellement deux ou trois boissons alcoolisées lorsqu'ils buvaient avaient environ 25 % plus de risque d'avoir fait l'essai du vapotage que les autres types de consommateurs d'alcool et que ceux ne buvant pas. Le risque d'avoir fait l'expérience du vapotage augmentait avec la taille de l'école dans la plage jusqu'à 500 élèves, et il demeurait élevé jusqu'à ce que la taille de l'école atteigne environ 1850 élèves. Le risque baissait très légèrement pour les écoles comptant entre 1850 et 2000 élèves.

En ce qui concerne le vapotage quotidien, nous avons constaté un risque accru de vapotage quotidien, par rapport aux répondants ne l'ayant pas fait, chez les répondants ayant déjà consommé du cannabis ou du tabac ou des boissons caféinées au cours des 12 derniers mois ou des 7 derniers jours (voir l'annexe à https://osf.io/x36p8). Pour la source de cigarette électronique, le risque qu'un répondant soit vapoteur quotidien était très différent selon s'il avait emprunté une cigarette électronique à un ami ou s'il en avait acheté une dans un commerce. Pour le type de cigarette électronique, le risque d'être vapoteur quotidien était de 25 % inférieur chez les répondants utilisant des cigarettes électroniques sans nicotine. Les répondants n'ayant jamais consommé de cannabis étaient légèrement moins à risque d'être vapoteurs quotidiens que les répondants ayant consommé du cannabis de diverses sources. Tout absentéisme scolaire au cours des 20 derniers jours était associé à un risque accru de vapotage quotidien et, bien qu'il soit possible que le vapotage quotidien ait entraîné une augmentation de l'absentéisme scolaire, notre modèle n'était pas conçu pour établir la preuve d'une telle relation.

Interactions

Les 10 corrélats principaux de l'essai de vapotage, à l'exception de l'état d'ivresse, offraient tous des interactions avec l'ethnicité (voir l'annexe à https://osf.io/x36p8). Le fait d'avoir déjà consommé des produits du tabac ou du cannabis et d'avoir consommé de l'alcool au cours des 12 derniers mois présentait des interactions avec l'ethnicité, le SSE et l'âge. Un répondant d'origine japonaise avait une probabilité plus élevée d'avoir vapoté au moins une fois qu'un individu d'origine non japonaise, quelle que soit la taille de l'école, tandis que des relations inverses étaient observées chez les répondants originaires d'Asie du Sud-Est et de Corée. Pour toutes les sources de cannabis, il était moins probable qu'un répondant d'origine non japonaise ait essayé le vapotage qu'un répondant d'origine japonaise. Quel que soit le groupe ethnique, un répondant qui consommait deux ou trois verres d'alcool au cours d'une journée typique avait la plus forte probabilité d'essayer le vapotage, comparativement aux autres types de consommation d'alcool. Alors que l'origine japonaise a été associée positivement à la probabilité d'avoir déjà vapoté, le fait d'être originaire d'Asie du Sud-Est ou de Corée était inversement associé à l'expérience de vapotage. Il y avait de plus petites différences dans la probabilité d'avoir fait l'expérience de vapotage pour un répondant d'origine japonaise et cette probabilité pour un répondant d'une origine autre que japonaise pour la consommation de cannabis ou d'alcool et la consommation d'alcool combinée à des boissons

énergisantes au cours des 12 derniers mois. Cette relation a également été constatée pour les répondants ayant consommé des produits du tabac ou du cannabis au cours de leur vie. Dans tous les groupes de SSE, être d'origine d'Asie du Sud-Est ou de Corée était associé à une probabilité légèrement plus faible d'avoir fait l'expérience du vapotage par rapport aux autres origines ethniques.

L'âge présentait des interactions avec la consommation d'alcool dans les 12 derniers mois, l'essai du tabagisme et l'essai de consommation de cannabis et, parmi ces interactions, la consommation d'une substance était un prédicteur plus important chez les jeunes élèves que chez les élèves plus âgés. Pareillement, ces variables étaient des prédicteurs plus importants chez les élèves ayant un SSE plus élevé que chez ceux ayant un SSE plus faible.

Une légère interaction a été observée pour le vapotage quotidien entre la consommation de caféine et l'ethnicité (voir l'annexe à https://osf.io/x36p8/). La force de l'interaction entre le fait d'avoir consommé des boissons caféinées au cours des 7 derniers jours et celui d'avoir une incertitude à propos de l'origine ethnique était de 0,111. Le fait d'avoir consommé des boissons caféinées au cours des 7 derniers jours était associé à une probabilité légèrement plus élevée de vapotage quotidien, indépendamment de l'incertitude quant à l'ethnicité.

Analyse de sensibilité

Conformément aux résultats de l'analyse primaire, une performance élevée a été observée dans les algorithmes de parcimonie de forêt aléatoire avec les 10 corrélats principaux seuls. Le modèle de parcimonie du vapotage quotidien a fourni une précision de 0,81 (IC à 95 % : 0,78 à 0,84), une sensibilité de 0,80 (0,72 à 0,86), une spécificité de 0,82 (0,78 à 0,85) et une ASC de 0,87, tandis que le modèle de parcimonie d'essai du vapotage a fourni une précision de 0,78 (0,76 à 0,79), une sensibilité de 0,78 (0,74 à 0,81), une spécificité de 0,78 (0,75 à 0,80) et une ASC de 0,86. À l'opposé, les régressions logistiques de base des deux résultats ont offert une performance inférieure à celle des modèles de forêt aléatoire de l'analyse primaire. Plus précisément, le modèle logit du vapotage quotidien a fourni une précision de 0,53 (0,49 à 0,57), une sensibilité de 0,63 (0,54 à 0,71), une spécificité de 0,50 (0,45 à 0,54) et une ASC de 0,60 tandis que le modèle logit de l'essai de vapotage a fourni une précision de 0,61 (0,59 à 0,64), une sensibilité de 0,82 (0,79 à 0,85), une spécificité de 0,52 (0,49 à 0,55) et une ASC de 0,73.

Analyse

Nous avons appliqué une approche d'apprentissage automatique pour étudier les corrélats du vapotage quotidien et de l'essai de vapotage à l'aide des données du SCDSEO recueillies auprès d'un échantillon représentatif de jeunes de l'Ontario fréquentant une école primaire ou secondaire. Les algorithmes finaux de forêt aléatoire ont fourni une performance élevée. Les 10 corrélats principaux du vapotage quotidien se sont révélés différents des 10 corrélats principaux de l'essai de vapotage, ce qui correspond aux divers prédicteurs connus dans la recherche sur le tabagisme[29-31] pour le début de l'utilisation de la cigarette et pour l'intensification de cette utilisation. Alors que nous n'avons relevé aucune interaction entre les paires de prédicteurs choisies pour le vapotage quotidien, nous avons relevé des interactions entre plusieurs prédicteurs de l'essai de vapotage, particulièrement en fonction de l'ethnicité.

Notre étude semble indiquer qu'il existe des différences entre les principaux corrélats de l'essai de vapotage et ceux du vapotage quotidien. Bien qu'une étude antérieure ait conclu que les influences sociales sont les prédicteurs les plus puissants d'une première expérience de vapotage[32], notre étude souligne l'importance de trois substances, à savoir le cannabis, l'alcool et le tabac, pour ce risque. Ces résultats concordent avec la tendance émergente au vapotage de cannabis[33] et montrent que la nicotine, un composé hautement toxicomanogène du tabac, est la substance la plus répandue dans les dispositifs de vapotage[34]. Nous avons également déterminé que la taille de l'école est un corrélat sociodémographique important du risque d'essai de vapotage.

Pour la source de cigarette électronique, puisque le risque le plus faible de vapotage quotidien a été observé chez les répondants ayant essayé la cigarette électronique d'un ami ou en ayant emprunté une, les influences sociales pourraient jouer un rôle limité dans le développement du vapotage quotidien. L'utilisation de cigarettes électroniques contenant de la nicotine a été associée au risque le plus élevé de vapotage quotidien, ce qui n'est pas surprenant, puisque la nicotine dicte la dépendance au vapotage[35]. Nos résultats montrent que la caféine, le cannabis et le tabac sont susceptibles d'accroître le risque de vapotage quotidien. Bien que la littérature indique que le niveau scolaire et l'âge pourraient être les corrélats sociodémographiques les plus puissants dans la consommation de drogues[36], notre étude montre qu'un absentéisme scolaire accru au cours des 20 derniers jours pourrait contribuer davantage à l'augmentation du risque de vapotage quotidien.

Points forts et limites

Sur le plan méthodologique, notre étude fournit de nouvelles données sur l'utilité de l'apprentissage automatique dans la conception de modèles prédictifs pour lutter contre le tabagisme[37]. La haute performance du modèle de forêt aléatoire fournit des résultats interprétables, comme l'identification de caractéristiques importantes, qui pourraient être utiles aux décideurs. Étant donné que les recherches indiquent que l'utilisation de la cigarette électronique à l'adolescence est associée à une probabilité accrue de fumer des cigarettes[38], les caractéristiques mises en exergue fournissent des corrélats importants susceptibles de favoriser l'aide aux jeunes afin qu'ils ne fassent pas la transition vers le tabagisme. L'absentéisme scolaire et la taille de l'école, des indicateurs rarement rencontrés dans la littérature, ont été identifiés comme des corrélats importants des résultats grâce à l'utilisation de cette méthodologie d'apprentissage automatique.

En outre, la performance élevée observée dans cette étude va dans le sens de la recherche, qui démontre que l'apprentissage automatique peut parfois surpasser la modélisation statistique conventionnelle. Par exemple, un examen systématique révèle que les modèles d'apprentissage automatique ont une performance plus élevée que celle de la régression logistique dans les prédictions de résultats neurochirurgicaux[39]. De même, les modèles d'apprentissage automatique présentent des indices de concordance plus élevés que les scores du risque clinique dans la performance pronostique chez les patients ayant un saignement gastro-intestinal aigu[40].

En ce qui concerne les limites, puisque notre étude était transversale, nous n'avons pas pu prouver que les 10 corrélats les plus importants étaient de véritables prédicteurs d'un vapotage quotidien ou de l'essai de vapotage. Malgré la rigueur des algorithmes de forêt aléatoire[41], l'importance relative des corrélats n'induit pas de causalité et nous n'avons pas effectué de test d'hypothèse dans cette analyse. Des études longitudinales futures comportant un modèle de recherche et une analyse axés sur la causalité contribueraient à combler cette limite. D'autres recherches sont également nécessaires pour valider les résultats au sujet des interactions, puisque les échantillons de groupes ethniques de l'étude étaient relativement petits (n < 150). Bien que nos modèles aient démontré un rendement élevé avec une simple imputation des données manquantes, il serait utile dans les recherches à venir d'envisager des méthodes plus sophistiquées comme l'imputation multiple si la précision des corrélats est privilégiée[42].

En outre, les outils dont on dispose pour élaborer des algorithmes de forêt aléatoire ne sont pas en mesure d'intégrer un échantillonnage par grappes. Toutefois, cette limitation affecte seulement la variance des corrélats, ce qui n'était pas l'objet de cette étude. Enfin, notre analyse comporte des limites inhérentes aux études d'enquête, comme de potentiels biais de rappel et de réponse. Néanmoins, nous nous attendons à ce que les résultats demeurent solides, car nous croyons que le sondage SCDSEO a été structuré avec des instruments qui optimisent la qualité des réponses.

Conclusion

En entraînant et en mettant à l'essai des algorithmes de forêt aléatoire, nous avons produit deux ensembles de 10 corrélats principaux différents pour le vapotage quotidien et pour l'essai de vapotage au sein d'une population de jeunes Canadiens. Nous avons observé des interactions entre certains corrélats importants et des caractéristiques sociodémographiques pour l'essai de vapotage. L'identification de ces corrélats dans un objectif de ciblage pour le vapotage quotidien et pour l'essai de vapotage va pouvoir certainement éclairer les futures études longitudinales visant à améliorer les politiques destinées à certains sous-groupes de population, indépendamment d'une relation de causalité.

Remerciements

Cette recherche est financée par les Instituts de recherche en santé du Canada (numéro de référence de financement MS2-17073).

Conflits d'intérêts

Les auteurs déclarent n'avoir aucun conflit d'intérêts.

Contributions des auteurs et avis

JS, HH et MC ont conceptualisé le manuscrit. JS a dirigé la rédaction, l'analyse statistique et l'interprétation des données, sous la direction de RF et de MC. Tous les auteurs ont fourni des commentaires, ont révisé les différentes versions du manuscrit et en ont approuvé la version définitive.

Le contenu de l'article et les points de vue qui y sont exprimés n'engagent que les auteurs; ils ne correspondent pas nécessairement à ceux du gouvernement du Canada.

Références

1. Hammond D, Reid JL, Rynard VL, et al. Prevalence of vaping and smoking among adolescents in Canada, England, and the United States: repeat national cross sectional surveys. BMJ. 2019;365:l2219. https://doi.org/10.1136/bmj.l2219

2. Stratton K, Kwan LY, Eaton DL, editors. Public health consequences of e-cigarettes. Washington (DC): National Academies Press (US); 2018. Chapter 8, Dependence and abuse liability; p. 255-338.

3. Vogel EA, Cho J, McConnell RS, Barrington-Trimis JL, Leventhal AM. Prevalence of electronic cigarette dependence among youth and its association with future use. JAMA Netw Open. 2020;3(2):e1921513. https://doi.org/10.1001/jamanetworkopen.2019.21513

4. Almeida-da-Silva CL, Matshik Dakafay H, O'Brien K, Montierth D, Xiao N, Ojcius DM. Effects of electronic cigarette aerosol exposure on oral and systemic health. Biomed J. 2021;44(3):252-259. https://doi.org/10.1016/j.bj.2020.07.003

5. Morean ME, Krishnan-Sarin S, O'Malley SS. Assessing nicotine dependence in adolescent e-cigarette users: the 4-item Patient-Reported Outcomes Measurement Information System (PROMIS) nicotine dependence item bank for electronic cigarettes. Drug Alcohol Depend. 2018;188:60-63. https://doi.org/10.1016/j.drugalcdep.2018.03.029

6. Foulds J, Veldheer S, Yingst J, et al. Development of a questionnaire for assessing dependence on electronic cigarettes among a large sample of ex-smoking e-cigarette users. Nicotine Tob Res. 2015;17(2):186-192. https://doi.org/10.1093/ntr/ntu204

7. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318. https://doi.org/10.1001/jama.2017.18391

8. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116-119. https://doi.org/10.1038/s41586-019-1390-1

9. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inf Decis Mak. 2018;18(Suppl 4):122. https://doi.org/10.1109/BIBM.2017.8217669

10. DuBrava S, Mardekian J, Sadosky A, et al. Using random forest models to identify correlates of a diabetic peripheral neuropathy diagnosis from electronic health record data. Pain Med. 2017;18(1):107-115. https://doi.org/10.1093/pm/pnw096

11. Caballero FF, Soulis G, Engchuan W, et al. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: The ATHLOS project. Sci Rep. 2017;743955. https://doi.org/10.1038/srep43955

12. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. https://doi.org/10.2196/jmir.5870

13. Fu R, Mitsakakis N, Chaiton M. A machine learning approach to identify correlates of current e-cigarette use in Canada. Explor Med. 2021; 2:74-85. https://doi.org/10.37349/emed.2021.00033

14. Choi J, Jung H-T, Ferrell A, Woo S, Haddad L. Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (hookah) users. J Clin Med. 2021; 10(5):972. https://doi.org/10.3390/jcm10050972

15. Fu R, Kundu A, Mitsakakis N, Chaiton M. Machine learning applications in tobacco research: a scoping review. Tob Control. Prépublication epub le 27 août 2021. https://doi.org/10.1136/tobaccocontrol-2020-056438

16. Park S, McCague H, Northrup D, Myles R, Chi T. The design and implementation of the CAMH Ontario Student Drug Use and Health Survey (OSDUHS) 2019: Technical documentation for Centre for Addiction and Mental Health. Toronto (Ont.) : Institute for Social Research, York University.

17. Kuhn M, Jed W, Steve W, et al. caret: classification and regression training. CRAN Repository; 2020. En ligne à : https://cran.r-project.org/web/packages/caret/caret.pdf

18. Breiman L. Random forests. Mach Learn. 2001;45:5-32. https://doi.org/10.1023/A:1010933404324

19. Srinath K. Ensemble machine learning: wisdom of the crowd [Internet]. Towards Data Science [plate-forme d'échange de données scientifiques]; 26 avril 2020 [consultation le 8 février 2021]. En ligne à : https://towardsdatascience.com/ensemble-machine-learning-wisdom-of-the-crowd-56df1c24e2f5

20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. https://arxiv.org/pdf/1106.1813.pdf

21. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law Hum Behav. 2005;29(5):615-620. https://doi.org/10.1007/s10979-005-6832-7

22. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Belmont (CA): Wadsworth International Group; 1984.

23. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. 2001;29(5): 1189-1232. https://doi.org/10.1214/aos/1013203451

24. Greenwell B. Partial dependence plots. CRAN Repository; 2018. En ligne à : https://cran.r-project.org/web/packages/pdp/pdp.pdf

25. Potter LN, Lam CY, Cinciripini PM, Wetter DW. Intersectionality and smoking cessation: exploring various approaches for understanding health inequities. Nicotine Tob Res. 2021; 23(1):115-123. https://doi.org/10.1093/ntr/ntaa052

26. Greenwell BM, Boehmke BC, McCarthy AJ. A simple and effective model-based variable importance measure. arXiv. 2018. https://arxiv.org/abs/1805.04755

27. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychology. 2000;19(6):586-592. https://doi.org/10.1037//0278-6133.19.6.586

28. Milborrow S. Plot a model's residuals, response, and partial dependence plots. 2020. En ligne à : https://cran.r-project.org/web/packages/plotmo/plotmo.pdf

29. Pokhrel P, Fagan P, Kawamoto CT, Okamoto SK, Herzog TA. Predictors of marijuana vaping onset and escalation among young adults. Drug Alcohol Depend. 2020;216:108320. https://doi.org/10.1016/j.drugalcdep.2020.108320

30. Wellman RJ, Dugas EN, Dutczak H, et al. Predictors of the onset of cigarette smoking: a systematic review of longitudinal population-based studies in youth. Am J Prev Med. 2016; 51(5):767-778. https://doi.org/10.1016/j.amepre.2016.04.003

31. Morean ME, Wedel AV. Vaping to lose weight: predictors of adult e-cigarette use for weight loss or control. Addict Behav. 2017;66:55-59. https://doi.org/10.1016/j.addbeh.2016.10.022

32. Jayakumar N, O'Connor S, Diemert L, Schwartz R. Predictors of e-cigarette initiation: findings from the Youth and Young Adult Panel Study. Tob Use Insights. 2020;13:1179173X20977486. https://doi.org/10.1177/1179173x20977486

33. Chadi N, Minato C, Stanwick R. Cannabis vaping: understanding the health risks of a rapidly emerging trend. Paediatr Child Health. 2020; 25(Suppl 1):S16-S20. https://doi.org/10.1093/pch/pxaa016

34. US Food and Drug Administration (FDA). Chemicals in tobacco products and your health—nicotine: the addictive chemical in tobacco products [Internet]. Washington (DC): FDA; 2020 [consultation le 15 mars 2021]. En ligne à : https://www.fda.gov/tobacco-products/health-effects-tobacco-use/chemicals-tobacco-products-and-your-health

35. Dinardo P, Rome E. Vaping: the new wave of nicotine addiction. Cleve Clin J Med. 2019;86(12):789-798. https://doi.org/10.3949/ccjm.86a.19118

36. Boak A, Elton-Marshall T, Mann RE, Hamilton HA. Drug use among Ontario students 1977-2019: detailed findings from the Ontario Student Drug Use and Health Survey (OSDUHS). Toronto (Ont.): Centre for Addiction and Mental Health; 2020. 312 p. En ligne à : https://www.camh.ca/-/media/files/pdf---osduhs/drugusereport_2019osduhs-pdf.pdf

?la = en&hash = 7F149240451E7421C3991121AEAD630F21B13784

37. Nam SJ, Kim HM, Kang T, Park CY. A study of machine learning models in predicting the intention of adolescents to smoke cigarettes. arXiv. 2019. https://arxiv.org/abs/1910.12748v2

38. Dutra LM, Glantz SA. Electronic cigarettes and conventional cigarette use among US adolescents: a cross-sectional study. JAMA Pediatr. 2014; 168(7):610-617. https://doi.org/10.1001/jamapediatrics.2013.5488

39. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg. 2018; 109:476-486.e1. https://doi.org/10.1016/j.wneu.2017.09.149

40. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. Dig Dis Sci. 2019; 64(8):2078-2087. https://doi.org/10.1007/s10620-019-05645-z

41. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci. 2017; 9:329. https://doi.org/10.3389/fnagi.2017.00329

42. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol. 2020. https://doi.org/10.1016/j.cjca.2020.11.010