AFCROs

LES ENTREPRISES DE LA RECHERCHE CLINIQUE

L'UTILISATION DE L'INTELLIGENCE ARTIFICIELLE (IA) DANS L'ANALYSE DES DONNÉES DE SANTÉ EN VIE RÉELLE

Rédacteurs pour l'AFCROs
Martin Prodel
Manon Belhassen
Camille Nevoret
Audrey Lajoinie
Pascale Rondeau

11/2022

1 LE CONTEXTE DES DONNÉES DE SANTÉ ET LA PLACE DE L'IA

Le monde de la recherche en santé dispose aujourd'hui de très nombreuses sources de données - au niveau national avec le Système National des Données de Santé (SNDS), ou plus localement via des registres, des établissements de santé ou des entrepôts de données - et doit faire face à la nécessité de développer des outils analytiques toujours plus adaptés et plus performants pour exploiter au mieux ces données. Créer ces outils analytiques est désormais un enjeu majeur, notamment pour la détection de signaux épidémiques, pour la pharmacovigilance, l'accès au marché des médicaments ou des dispositifs médicaux, ou encore le développement d'outils d'aide à la décision destinés à la pratique médicale. Cette démarche est au croisement de l'épidémiologie telle qu'elle existe depuis des décennies et des méthodes innovantes dites d'Intelligence Artificielle (IA). Au-delà de l'image parfois médiatique de l'IA, il est important de cerner comment ces techniques apportent en pratique de nouvelles réponses au regard de l'épidémiologie plus classique et établie, et les nouvelles perspectives qu'elles ouvrent. C'est avant tout une façon différente d'utiliser des données, notamment si elles sont volumineuses.

Qu'entend-on par "IA appliquée aux données de santé" ?

Quand on parle d'analyse des données de santé en vie réelle, l'IA se présente sous la forme et le nom d'apprentissage automatisé (en anglais : *machine learning*). Il s'agit en pratique d'une boîte à outils à disposition de l'analyste (aussi appelé *data scientist* ou statisticien), pour extraire et traiter la donnée. Chaque outil est un algorithme dont les plus connus sont les «forêts aléatoires» et les «réseaux de neurones». L'outil doit être minutieusement choisi, puis façonné sur mesure pour répondre à la problématique rencontrée : étude de mortalité, efficacité d'un traitement, analyse des parcours, quantification des risques de ré-hospitalisation, prédiction des patients à risque, etc. L'essor de l'IA est intimement lié à celui des grandes bases de données nationales (ex. : le SNDS), des entrepôts de données et des registres, car avoir suffisamment de données de qualité est un prérequis incontournable de l'IA.

L'IA : les mêmes prérequis qu'une approche conventionnelle

Le recours à l'IA ne doit pas être systématique pour analyser des données. L'IA n'est qu'un outil parmi d'autres et dont l'usage est réservé aux problématiques qui s'y prêtent (analyse automatique d'images ou de comptes-rendus, prédiction à partir de grandes bases de données…). Toutefois, l'IA est bel et bien là, opérationnelle et utile. Elle est déjà utilisée en pratique, par exemple pour la détection automatique d'anomalies sur les images médicales, mais aussi via des techniques de machine learning lors d'études épidémiologiques. Il s'agit d'IA conçues sur mesure pour répondre à une hypothèse spécifique. Pour cela, une question de recherche doit être clairement explicitée. Cela répond aux mêmes exigences méthodologiques que d'autres outils statistiques.

2 ABORDER DES THÉMATIQUES CONCRÈTES

Thématique n° 1 : L'IA au service de la médecine personnalisée - l'apport du *machine learning* dans l'aide à la décision en santé

La prédiction individuelle de la survenue d'un événement clinique (ex., complication, décès, succès d'une prise en charge) est un enjeu clé de la médecine personnalisée. Elle permet d'estimer un « risque » - une probabilité - de survenue pour un patient donné en fonction de ses caractéristiques, de son mode vie, de ses antécédents, etc. Quand ce risque aide le soignant dans le choix d'actions préventives ou curatives, on parle d'outil d'aide à la décision.

Un algorithme prédictif ou pronostic repose sur un modèle mathématique capable d'apprendre des caractéristiques et de l'historique des patients. Le modèle est entraîné, validé, puis testé en séparant les patients en plusieurs échantillons. Le modèle fera idéalement l'objet d'une validation externe, c'est-à-dire sur une autre cohorte de patients que celle utilisée pour son développement. Enfin, lorsque le modèle est destiné à être utilisé comme outil d'aide à la décision en pratique courante, des études d'impact sur la qualité des pratiques, les résultats de santé et la sécurité des soins doivent confirmer sa capacité à améliorer la santé des patients ; un essai randomisé apportera alors le meilleur niveau de preuve [1,2].

Historiquement, ces algorithmes étaient basés sur des méthodes statistiques «conventionnelles» (e.g. régression logistique multivariée). Si l'apport de ces modèles reste indiscutable, la capacité à générer des quantités de données de santé

de plus en plus importantes a abouti à l'utilisation de nouvelles méthodes, telles que le *Data Mining* (exploitation de données) et, tout particulièrement le *Machine Learning* (ML). Les méthodes de ML plus classiquement utilisées pour le développement d'outils prédictifs (modèles supervisés) sont le *support-vector machine* (SVM), le *gradient-boosting machine* (GBM), le *random forest* (RF) ou encore les réseaux de neurones [3]. Il est recommandé de tester plusieurs modèles afin de sélectionner le plus performant en appliquant une méthode de validation croisée (*cross-validation*) [4].

Lo-Ciganic *et al.* ont développé et validé un algorithme prédictif du risque d'overdose chez 14 000 patients traités par opioïdes, avec 284 facteurs de risque potentiels testés, à partir des données Medicaid (base US de remboursements de soins) [5]. Cet algorithme est destiné à identifier les patients les plus susceptibles de bénéficier du programme de prévention des surdosages. Plusieurs méthodes ont été testées ; le GBM a montré les meilleures performances. Une approche particulièrement intéressante était l'analyse dynamique des consommations de soins au cours du temps, qui permettait d'identifier non seulement les patients les plus susceptibles de bénéficier du programme de prévention, mais aussi le moment le plus opportun pour mettre en place ce programme.

A l'instar des méthodes statistiques conventionnelles, les méthodes de ML présentent des limites, notamment le risque de surapprentissage : le modèle devient trop dépendant des données d'entraînement et n'est plus capable de généraliser à de nouvelles données ; il ne pourra alors pas être utilisé en pratique. Par ailleurs, l'exemple ci-dessus aborde un cas concret autour de la survenue d'un événement clinique, une variable binaire. Les algorithmes peuvent aussi prédire des variables quantitatives, comme les coûts de prise en charge [6], par des méthodes dédiées (e.g., régression linéaire, réseaux de neurones, RF ou SVM).

Dans le cas des outils d'aide à la décision se posera aussi la question du choix et du développement de l'interface numérique ergonomique qui permettra l'utilisation de l'algorithme et la visualisation des prédictions dans un contexte de soins. Également, l'explicabilité des prédictions, c'est-à-dire l'apport d'informations pour étayer la raison d'une prédiction plutôt qu'une autre, est une étape clé pour l'acceptation et l'utilisation en pratique de l'outil.

Thématique n° 2 : le *clustering* de lignes de traitement

La modélisation des schémas de traitement permet d'identifier des profils de patients selon leurs consommations de soins ou à l'inverse d'identifier des schémas de traitement (modalités : doses, fréquence ; et séquences) spécifiques à un sous-groupe de patients, dans le but d'optimiser leur prise en charge [7].

Plusieurs éléments définissent un schéma de traitement et peuvent être privilégiés lors de leur analyse. Cette dernière peut se focaliser sur l'aspect séquentiel ou temporel de la succession des traitements, ou sur des intervalles de temps prédéfinis, sur le détail des traitements ou sur de grandes catégories, sur des informations binaires (utilisation ou non d'un traitement) versus des informations continues (nombre de délivrances ou posologie). Le choix de ces éléments est primordial car il détermine la méthode de modélisation des parcours et de représentation graphique adéquate (diagramme sankey, sunburst, tapis de séquences).

Afin d'identifier des sous-groupes de patients ayant des schémas de traitement proches, plusieurs algorithmes de clustering peuvent être utilisés, qu'ils soient hiérarchiques (CAH) ou non hiérarchiques (K-means, K-medoids). Une distance entre les trajectoires modélisées de chaque patient est calculée au préalable pour pouvoir appliquer les algorithmes. La distance de Hamming ou la distance basée sur la plus longue sous-séquence commune peuvent être utilisées. Le nombre de clusters est choisi en fonction de critères statistiques de qualité des partitions obtenues (silhouette, etc.) et d'un avis médical. Des analyses de sensibilité sont recommandées afin d'évaluer l'impact du nombre de clusters choisi.

Ces méthodes appliquées aux schémas de traitements peuvent être étendues à l'analyse de parcours de soins, ce qui permet de prendre en compte des consommations de soins de natures différentes (hospitalisations, actes thérapeutiques ou consultations). Elles peuvent aussi permettre de projeter la séquence de traitement d'un nouveau patient.

Thématique n°3 : Comment obtenir plus de données pour faire de l'IA ?

Les approches utilisant l'IA nécessitent l'usage d'un volume conséquent de données, et d'une donnée de qualité. La collecte de données étant un processus long et coûteux, il y a donc un fort enjeu à pouvoir maximiser l'usage de données déjà existantes. Toutefois, ce n'est pas sans présenter des difficultés, à la fois pour la collecte et pour le partage des données, face auxquelles il existe désormais des solutions technologiques.

Quand les données sont incomplètes, c'est-à-dire partiellement labellisées (images, dossier patient), il est possible d'utiliser une méthode mixte d'apprentissage machine pour analyser les données : l'apprentissage semi-supervisé. Plutôt que d'ignorer les données non labellisées, ou de ne mettre en place qu'une approche non supervisée (clustering) sur toutes les données, ces méthodes combinent les deux approches. Cela permet d'améliorer les modélisations, ou encore d'extrapoler une information disponible sur un sous-groupe (notamment via un retour au dossier patient ou dans un registre sur un sous-échantillon) vers une cohorte entière. L'apprentissage semi-supervisé est particulièrement adapté en bio-informatique pour le profilage de l'expression génique, ou en détection du cancer avec tomodensitométrie. Ces approches sont d'autant plus pertinentes pour le cas de très grosses masses de données qui ne sont pas toujours labellisables, ou qui nécessitent le développement de plateformes de labellisation dédiées.

Lorsque l'incomplétude des données en vient à limiter les usages, par exemple si cela empêche de comparer des groupes de patients ou de prédire des interactions complexes, on peut alors envisager de générer artificiellement des données supplémentaires. Ces données synthétiques vont rendre possibles des comparaisons jusqu'alors impossibles (ex.: essai clinique monobras, étude en vie réelle non contrôlée [8]). La génération de patients virtuels repose sur des caractéristiques probables au regard d'une base de données existante (les patients générés ressemblent à des patients réels), ou peut être purement hypothétique (les patients générés sont construits sur la base d'hypothèses scientifiques).

Un cas d'usage des données synthétiques est de créer des groupes témoins pour les essais cliniques pour lesquels on manque de données comparatives (maladies rares, pathologies émergentes). Une autre approche consiste à produire de la donnée pour améliorer la précision d'un modèle de machine learning : cela permet d'augmenter la quantité d'information

disponible, via un mix de données réelles et virtuelles. Quelles méthodes permettent de générer des données synthétiques ? Les approches les plus utilisées pour la reproduction d'une donnée existante, souvent multidimensionnelle et avec des relations complexes, sont les *Agent-based modelling* (ABM) [9] et les approches de *deep learning* comme les *Variational Auto-Encoders* (VAE) et les *Generative Adversarial Networks* (GANs) [10]. Ce sujet des données générées est en plein essor et sera très présent dans les toutes prochaines années, comme illustré par une étude de Gartner qui prévoit qu'en 2030, il y aura plus de données virtuelles que de données réelles [11].

Enfin, toujours dans cette idée de maximiser l'usage de données déjà existantes, une autre démarche consiste à utiliser la donnée là où elle se trouve sans la déplacer. C'est ce à quoi se consacre le *federating learning*. Avec cette technique d'analyse de machine learning, il n'y a plus besoin de centraliser toutes les données de santé au même endroit pour y appliquer l'analyse, comme c'est le cas lorsque les données proviennent de plusieurs établissements de santé. C'est l'algorithme qui est morcelé et envoyé à chaque endroit où les données sont présentes (c'est-à-dire chez chaque propriétaire). Ensuite, chaque morceau fragmenté localement de l'algorithme va modéliser les données qui sont à sa disposition, pour être *in fine* ré-agrégé en un seul modèle de façon centralisée. Le modèle complet ainsi obtenu est utilisable et disponible pour chacun des acteurs ayant participé à sa conception. Cette approche résout des problèmes de sécurité et de confidentialité des données de santé, car celles-ci n'ont plus à être déplacées et/ou partagées en dehors de leur lieu de stockage. Un challenge fort pour la mise en place opérationnelle du *federated learning* est une contrainte technique : il requiert l'interopérabilité des données (format des données, dictionnaire commun de variables, structure) entre les différentes sources de données pour fonctionner [12].

LE MOT DE LA FIN

La richesse des bases de données de santé permet l'émergence de modèles prédictifs sur l'amélioration des parcours de santé et sur une personnalisation des traitements. Les méthodes de machine learning sont d'ores et déjà bien implantées, comme illustré par les nombreuses publications scientifiques de ces dernières années. Ces approches sont également décrites et présentes dans les travaux des autorités de santé. Le guide de la HAS sur les études en vie réelle pour l'évaluation des médicaments et dispositifs médicaux fournit une description détaillée de ces méthodes, de leurs limites et des attentes qu'elles suscitent [13].

Au-delà de l'utilisation de l'IA dans les études épidémiologiques, les algorithmes sont en effet déjà utilisés et embarqués dans des objets connectés et des dispositifs médicaux. Cela pose de nouvelles problématiques concrètes de validation clinique des algorithmes de machine learning dans le cadre du soin, par exemple via de nouveaux designs d'essais cliniques et d'études sur données secondaires.

Références

1. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. British Medical Journal Publishing Group; 2016;353:i3140.

2. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691-8.

3. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Medical Research Methodology. 2019;19:64.

4. Berrar D. Cross-Validation. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, éditeurs. Encyclopedia of Bioinformatics and Computational Biology [Internet]. Oxford: Academic Press; 2019 [cité 23 août 2022]. p. 542-5. https://www.sciencedirect.com/science/article/pii/B978012809633820349X

5. Lo-Ciganic W-H, Donohue JM, Yang Q, Huang JL, Chang C-Y, Weiss JC, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. The Lancet Digital Health. Elsevier; 2022;4:e455-65.

6. Taloba AI, Abd El-Aziz RM, Alshanbari HM, El-Bagoury A-AH. Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning. Journal of Healthcare Engineering. Hindawi; 2022;2022:e7969220.

7. Roux J, Grimaud O, Leray E. Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis. Statistical Methods in Medical Research, SAGE Publications, 2019, 28 (6), pp.1651-1663. 10.1177/0962280218772068. hal-01798652.

8. https://www.afcros.com/wp-content/uploads/2022/07/V3-Compte-rendu-colloque-AFCROs-1.pdf

9. Andrew Crooks, Alison Heppenstall, Nick Malleson, 1.16 - Agent-Based Modeling, Editor(s): Bo Huang, Comprehensive Geographic Information Systems, Elsevier, 2018, Pages 218-243, ISBN 9780128047934,

10. Saxena, Divya & Cao, Jiannong. (2020). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions.

11. https://research.aimultiple.com/synthetic-data/

12. https://tripleblind.ai/article/how-is-federated-learning-used-in-healthcare

13. HAS (2021) Guide méthodologique : Études en vie réelle pour l'évaluation des médicaments et dispositifs médicaux, 10/06/2021. https://www.has-sante.fr/jcms/p_3284524/fr/etudes-en-vie-reelle-pour-l-evaluation-des-medicaments-et-dispositifs-medicaux