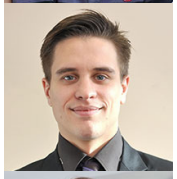
Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé

Philippe BESSE[1]

Université de Toulouse – INSA, Institut de Mathématiques de Toulouse – UMR CNRS 5219 – Chercheur régulier à l'ObvIA[2]



Aurèle BESSE-PATIN

McGill University & Montreal Neurological Institute

Céline CASTETS-RENARD

Université d'Ottawa – Titulaire de la chaire Law, Accountability and Social trust in AI, ANITI[3] – Chercheure régulière à l'ObvIA

TITLE

Legal and Ethical Implications of Artificial Intelligence Algorithms in the Health Field

RÉSUMÉ

L'Intelligence Artificielle (IA) envahit nos quotidiens et le domaine de la santé notamment pour aider au diagnostic, faire des choix thérapeutiques ou encore viser une médecine prédictive de précision. Absente de la loi française de bioéthique du 7 juillet 2011, l'IA fut très présente lors des États Généraux accompagnant la révision de la loi en 2018. La profusion de guides ou recommandations éthiques sur l'IA (*soft law*), motivés par la nécessité de conquérir la confiance des usagers, incite préalablement à se préoccuper de leur vigueur normative, en lien avec les textes juridiques promulgués depuis l'entrée en vigueur le 25 mai 2018 du RGPD (règlement 2016/679/UE – règlement général de protection des données personnelles). Une analyse conjointe de ces textes, des algorithmes d'IA déployés et d'applications concrètes en santé permet de poser les principales questions éthiques et légales soulevées dans ce domaine : principe du *consentement libre et éclairé* du patient face à l'opacité des algorithmes, risques potentiels de *discrimination* dans l'accès au soin, *intérêt public* ou *bien commun* attendu de la recherche en comparaison des *risques* encourus par l'ouverture de l'accès aux données personnelles. Les réponses conduisent à des recommandations déontologiques ou règlementaires indispensables à la transparence de ces outils : *protection* drastique des données de santé, notamment génétiques, et de leurs utilisations, rigueur des pratiques de recherche pour produire des *résultats reproductibles* donc scientifiques, *détection des biais* avant certification des dispositifs de santé et explicitation du *protocole d'information* des patients.

Mots-clés **:** *intelligence artificielle, apprentissage automatique, statistique, RGPD, code de santé publique, éthique, bioéthique, discrimination, droit de l'IA.*

ABSTRACT

Artificial Intelligence (AI) is invading our daily lives and the health field, notably to help with diagnosis, to make therapeutic choices or even to aim for precise predictive medicine. Absent from the French bioethics law of July 7, 2011, AI was very present during the "États Généraux" accompanying the revision of the law in 2018. The profusion of ethical guides or recommendations on AI (soft law), motivated by the need to win the trust of users,

1. philippe.besse@math.univ-toulouse.fr

2. ObvIA : Observatoire international sur les impacts sociétaux de l'IA et du numérique, https://observatoire-ia.ulaval.ca/
3. ANITI : *Artificial and Natural Intelligence Toulouse Institute*, https://aniti.univ-toulouse.fr/

encourages us to be concerned about their normative force, in connection with the legal texts promulgated since the entry into action on 25 May 2018 of the GDPR (regulation 2016/679/EU – general regulation on the protection of personal data). A joint analysis of these texts, of the AI algorithms deployed and of concrete applications in health, enables us to consider the main ethical and legal questions raised in this field: the principle of *free and informed consent* of the patient faced to the opacity of algorithms, potential risks of *discrimination* in access to care, *public interest* or *common good* expected from research in comparison with *risks* incurred by opening access to personal data. The responses lead to ethical or regulatory recommendations that are essential for the transparency of these tools: drastic *protection* of health data, particularly genetic data, and their uses, rigorous research practices to produce *reproducible* and therefore scientific *results*, *detection of biases* before certification of health devices and clarification of the patient *information protocol*.

***Keywords*:** *artificial intelligence, machine learning, statistics, GDPR, public health code, ethics, bioethics, discrimination, AI law.*

1. Introduction

1.1 Battage médiatique

L'intelligence artificielle (IA) dite *faible*, opposée à une IA *forte* supposée disposer d'une conscience de soi et que nous laisserons à la science-fiction, recouvre une grande variété d'objets, méthodes *et al*gorithmes susceptibles d'imiter des comportements humains « intelligents » : robots, véhicules autonomes, systèmes experts, algorithmes d'apprentissage automatique…

Depuis 2012, nous sommes soumis à une déferlante médiatique sans précédent sur les applications des algorithmes d'IA associées à des succès retentissants : reconnaissance d'images et diagnostic automatique, véhicules autonomes, victoire au go, traduction automatique… Ce battage médiatique fait suite à celui sur l'avènement du stockage tous azimuts de données massives ou *big data* et leur utilisation pour alimenter les nouveaux algorithmes d'IA exécutés dans des environnements technologiques en constante progression. Cette convergence entre données massives, algorithmes performants et puissance de calcul est à l'origine de l'expansion exceptionnelle des usages de l'IA dans tous les domaines de nos quotidiens. Les principaux acteurs technologiques comme *Google*, *Facebook*, *Amazon* ou *Microsoft*, ont tout intérêt à sur-médiatiser ces succès puisque leurs considérables revenus proviennent de la vente de l'application de ces technologies à notre profilage publicitaire. Ils se doivent donc d'en promouvoir l'efficacité, même si ses succès diffèrent en fonction du domaine d'application et si elle peut s'avérer anxiogène dans ses conséquences sociétales, tant sur la destruction d'emplois même qualifiés, que sur la déresponsabilisation des acteurs humains ou encore l'exposition des données de la vie privée.

1.2 Confiance et acceptabilité

Une composante importante de la publicité excessive autour de l'IA concerne son *acceptabilité*, comme celle de toute nouvelle technologie pénétrant ou plutôt envahissant nos quotidiens. Le principal enjeu est de cultiver ou conquérir la confiance des utilisateurs, qu'ils soient consommateurs, clients, patients, contribuables, justiciables ou citoyens, pour une IA acceptable. En première ligne, les entreprises privées spécialistes des réseaux sociaux et technologies numériques, rejointes ensuite par plus de 90 partenaires, se sont empressées, dès 2015, de signer une Charte de partenariat[4] pour une *IA au bénéfice du peuple et de la société*. Dès lors, tous les acteurs publics institutionnels ont rejoint le mouvement ; citons parmi les plus récents la partie 5 du rapport Villani pour *donner un sens à l'IA* (Villani *et al.*, 2018), les lignes directrices pour une *IA digne de confiance* des hauts experts désignés par la Commission Européenne (High Level Expert Group, 2019), ou encore la *déclaration de Montréal pour le développement d'une IA responsable* (Université de Montréal, 2018). Notons la création (12/2019) du Comité Pilote d'Éthique du Numérique[5] sous l'égide du Comité Consultatif National d'Éthique pour les Sciences de la Vie et de la Santé et dont le rapport attendu début 2021 doit s'intéresser aux liens entre diagnostic médical et IA. C'est plus largement une avalanche de recommandations pour une IA éthique au service de l'humanité dont Fjeld *et al.* (2019) proposent une analyse graphique et sémantique tandis que Jobin *et al.* (2019) en explore le paysage. Les enjeux sont considérables car, en l'absence de confiance, les utilisateurs n'accepteront pas l'IA. Sans acceptation sociale, les entreprises technologiques ne pourront plus collecter toutes les données nécessaires et ne pourront pas développer une IA pertinente, source de profits. Les conséquences de l'affaire *Cambridge Analytica* sur l'encours boursier de *Facebook*, en mars 2018, en furent une démonstration éclatante (Guichard, 2018).

4. https://www.partnershiponai.org/tenets/
5. https://www.ccne-ethique.fr/sites/default/files/communique_lancement_comite_numerique.pdf

1.3 Éthique et protection juridique

Cette affaire peut être citée parmi d'autres : condamnations successives de *Google* pour entrave à la concurrence, fuites massives et répétées de données personnelles, utilisations abusives de celles-ci… nous rappellent que le but premier des entreprises commerciales ou de leurs dirigeants n'est pas l'altruisme ou la philanthropie mais des encours boursiers et le montant des dividendes à distribuer à leurs actionnaires. Ces profits nécessitent des pratiques éthiques pour être acceptables mais la confiance des usagers sera nettement plus franche et massive si elle repose sur une protection juridique, plutôt que sur de bonnes intentions éthiques (*ethical washing*), aussi louables soient-elles. En France, la première version de la loi Informatique et Liberté date de 1978. Ce texte précurseur marqua une réelle anticipation des problèmes à venir. En revanche, à l'heure actuelle, la loi peine à suivre les évolutions ou disruptions technologiques. Ce sont bien entre autres quelques-uns de ces retards que vise à combler une révision de la loi de Bioéthique.

L'entrée en vigueur du RGPD (Commission Européenne, 2018), puis son intégration dans les textes nationaux des États membres, signent une avancée majeure pour la protection des données personnelles en Europe. Le principe de sécurité et confidentialité, au cœur de l'action de la Commission Nationale de l'Informatique et des Libertés (CNIL) en France, est en effet une priorité mais d'autres aspects, tant juridiques qu'éthiques, sont à considérer pour instaurer ou restaurer la confiance des usagers envers ces nouvelles technologies. Ainsi, l'article 22§1 du RGPD (Commission Européenne, 2018) accorde aux personnes concernées le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, produisant des effets juridiques la concernant ou l'affectant de manière significative. Repris dans les lois nationales des États membres, cet article a pu servir de fondement en droit français pour reconnaître un droit à l'*explicabilité* des décisions algorithmiques, dans le souci de lutter contre les risques de *discrimination*. Ces préoccupations rejoignent les exigences publiques exprimées dans un sondage réalisé au Royaume-Uni (Vayena *et al.*, 2018) au sujet des applications de l'IA en médecine.

Un large consensus est donc établi sur la nécessité de pratiques en IA respectueuses de l'éthique. Néanmoins, compte tenu des pressions financières, un cadre juridique s'avère indispensable. Il est un préalable à des pratiques vertueuses génératrices de confiance.

Tel est bien l'objectif de la Commission Européenne (CE) qui propose *les éléments clefs d'un futur cadre réglementaire* dans le livre blanc (Commission Européenne, 2020) *pour une IA basée sur l'excellence et la confiance fondée sur les droits fondamentaux de la dignité humaine et la protection de la vie privée*. La rédaction de ce livre blanc s'appuie sur les lignes directrices pour une IA de confiance (High Level Expert Group, 2019) rédigées par un groupe d'experts et dont il est important d'anticiper l'impact à venir. En résumé, les technologies de l'IA se développent à grande vitesse dans un contexte juridique très complexe mais insuffisant à encadrer les risques sociaux susceptibles de se produire. Ce cadre légal est appelé à évoluer, au moins en Europe, afin de minimiser les risques et créer les conditions d'une acceptabilité sociale de l'IA. Les limites de la norme légale étant identifiées, il est alors possible, dans un deuxième temps, de compléter la norme légale par des chartes de déontologie, telle celle des professionnels de la statistique publique européenne (Eurostat, 2017). Rappelons que l'objectif de ces chartes est tout autant de constituer une obligation réglementaire envers les employés qu'une protection de leur rigueur professionnelle contre les pressions extérieures politiques ou financières.

1.4 IA et bioéthique

Le champ d'étude ainsi esquissé est trop vaste. Nous proposons d'en limiter le périmètre en considérant le **domaine d'application restreint mais très sensible de la santé**, donc de la

bioéthique. Les questions de sécurité et confidentialité des données personnelles sont déjà abondamment traitées par des systèmes d'analyse (*Privacy Impact Assessment*[6] ou études d'impact des données personnelles) mis en place par la CNIL et rendus obligatoires par le RGPD en présence de risques qui peuvent notamment résulter du traitement de données sensibles, telles les données de santé (art. 35). Nous nous focaliserons sur *les questions touchant au risque de discrimination* et à la nécessité d'une explication intelligible des décisions, en lien avec les réglementations et textes juridiques concernés. D'autres auteurs (Racine *et al.*, 2019 ; Wiens *et al.*, 2019) ont récemment abordé ce sujet mais en privilégiant le point de vue médical, ainsi que ses interactions avec les nouvelles technologies et certaines questions éthiques émergentes. Aborder la question à partir du cadre juridique ouvre une autre perspective. Des questions essentielles sont alors soulevées, tenant en particulier à la notion de *consentement libre et éclairé* des personnes et aux *risques de discriminations*. D'autres questions émergent, qui nécessitent des réponses plus délicates à formuler, sur l'équilibre entre, d'une part, le développement de la recherche en santé et *l'intérêt public* ou *bien commun* attendu et, d'autre part, sur les risques afférents à l'accessibilité des données personnelles de santé.

Aborder la bioéthique des applications d'intelligence artificielle dans le domaine de la santé mobilise les compétences de nombreuses disciplines. L'objectif est pour le moins ambitieux. Aussi, pour en faciliter la lecture, cet article introduit de façon pédagogique les prérequis de chaque discipline, dans une volonté de faciliter les échanges réciproques. La section 2 définit plus précisément l'IA considérée dans cet article. La section 3 décrit brièvement les règles légales applicables à l'IA en santé. Les principaux domaines de santé concernés par ces modèles *et al*gorithmes sont cernés en section 4. La section 5 met en relation les champs disciplinaires du droit, de la santé, des sciences du numérique et de la statistique pour en tirer les conséquences : quelles sont les protections juridiques existantes ? Où sont leurs limites ou insuffisances qui nécessiteraient plus de réglementation ou un code déontologique des professionnels concernés ?

Les questions de bioéthique sont très culturellement marquées ainsi donc que les lois qui sont très spécifiques à un pays, notamment en France où elles s'avèrent plus restrictives par exemple sur l'accès aux tests génétiques par rapport à d'autres pays européens. Cet article, axé sur le corpus juridique européen et plus particulièrement français, ne peut évidemment prétendre à l'exhaustivité. En revanche, il peut constituer une première étape pour aborder de façon comparative les différentes situations.

Cette réflexion conduit, en conclusion, à proposer un ensemble de recommandations résumées dans le Tableau 1 permettant de préciser les responsabilités et devoirs de chaque partie : développeur, médecin, chercheur, patient, dans les actes médicaux aidés par l'IA ou pour la poursuite de recherches en santé sur des grandes bases de données personnelles.

6. https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil

Tableau 1 – *Proposition de recommandations à trois niveaux : 1. accès aux données ; 2. déontologie de la recherche ; 3. réglementation des dispositifs de santé intégrant de l'IA*

1. L'accès aux données nationales de santé peut être ouvert (INDS, CEIP) sans consentement explicite des personnes concernées lorsque les résultats attendus le justifient : cohortes épidémiologiques, diagnostic par imagerie médicale ou protéomique, étude des maladies rares ou monogéniques… En revanche, au regard des risques encourus de ré-identification, l'accès aux données de santé publique n'est pas justifié pour des projets de recherche pangénomiques sur les maladies multifactorielles.

2. Compte tenu des enjeux et des risques encourus, les équipes de recherche doivent se soumettre à un audit externe effectif et pas seulement déclaratif de sécurité de la chaîne d'archivage et de traitements des données personnelles de santé même pseudonymisées. Elles doivent s'astreindre à une rigueur d'analyse (détection, correction des biais) et d'évaluation des erreurs, afin de publier des résultats reproductibles, première exigence vers une certification. Elles doivent donner accès aux séquences de traitement des algorithmes lors d'une soumission avant publication.

3. Les autorités de santé (*e.g.* HAS, FDA…) ont la responsabilité de la certification ou du remboursement des dispositifs de santé intégrant de l'IA. Il importe d'*harmoniser* leurs protocoles en anticipant la stratégie en cours d'élaboration de la Commission Européenne. Obligation au responsable d'un système d'IA de produire une *documentation exhaustive* décrivant comment sont : (i) validées qualité, robustesse, résilience, des décisions ; (ii) traqués les biais des données ; (iii) intégré un suivi qualité adaptatif par enrichissement des bases d'apprentissage ; (iv) identifiés les responsables à chaque étape des traitements (recueil des données, entraînement des algorithmes, validation, certification, exploitation) pour la mise en place d'une boucle vertueuse de rétroaction. Ces autorités doivent formaliser les protocoles d'explicitation auprès des patients du rôle des algorithmes dans leur prise en charge, des risques d'erreur dans l'aide à la décision et des risques encourus de leur ré-identification.

2. De quelle IA est-il question ?

2.1 Historique

L'IA est apparue sous cette appellation dès 1955 à la suite du développement des premiers ordinateurs et a commencé à être formalisée par les travaux pionniers d'Alan Turing. La notion de neurone formel est due à McCulloch (neurophysiologiste) et Pitts (logicien) en 1943 tandis que le premier réseau de neurones est proposé par Rosenblatt (1958) avec un *perceptron* censé simuler le fonctionnement de la rétine. Faute de méthodes et capacités de calcul suffisantes, cette approche de l'IA a été mise en veilleuse au profit des *systèmes experts* dans les années 70. Ces systèmes associent une base de règles logiques, explicitées par des experts humains du domaine d'application, une base de faits et un moteur d'inférence. Ce dernier met itérativement en relation faits et prémices des règles pour en déduire de nouveaux faits jusqu'à atteindre le ou les faits correspondant à la décision recherchée ou l'objectif visé. Un tel prototype de système expert (Mycin) a été développé par Buchanan et Shortliffe (1984) pour la sélection d'un antibiotique adapté aux paramètres biologiques du diagnostic d'une infection bactérienne.

Malgré les très grandes difficultés de construction des bases de règles expertes, leur manque de flexibilité, ainsi que la complexité algorithmique exponentielle de leur exécution, cette approche n'a pas été complètement abandonnée (Darlington, 2011). Néanmoins la recherche sur les systèmes experts passa en arrière-plan à la fin des années soixante-dix au profit d'un retour des réseaux de neurones bénéficiant de moyens de calculs suffisants et de résultats théoriques (Rumelhart *et al.*, 1986) sur la convergence (locale) de l'algorithme de *rétropropagation du gradient* permettant d'entraîner itérativement un réseau multicouches. Dans un réseau, la connaissance est dite répartie, dans les valeurs des poids des entrées des neurones appris sur les données, par opposition à la connaissance localisée des bases de règles construites par les experts ; une IA *empirique* opaque s'oppose à une IA *symbolique* explicable. Les années quatre-vingts ont connu un développement massif de différents types de réseaux de neurones *et al*gorithmes d'apprentissage parallèlement à l'extension des méthodes et modèles statistiques appliqués à des objets complexes de grande dimension (courbes et fonctions).

Dans les années quatre-vingt-dix, ces réseaux se trouvèrent en concurrence avec bien d'autres algorithmes : modèles statistiques avec pénalisation, *k* plus proches voisins, arbres binaires de décision, machine à vecteurs supports, *boosting*, forêts aléatoires… (James *et al.*, 2013), poursuivant les mêmes objectifs prédictifs au sein d'une très large communauté scientifique réunie autour de l'apprentissage automatique (*machine learning* ou *ML*) à l'interface entre Sciences du Numérique, Mathématiques et Statistique. Les recherches sur les réseaux de neurones ont toujours progressé jusqu'à leur succès retentissant en 2012 sous l'appellation très médiatisée d'apprentissage profond (*deep learning*). Ces réseaux associent des dizaines de couches de neurones dont celles dites *convolutionnelles* ou d'autres *récurrentes* (*LSTM*) qui firent franchirent des étapes décisives, par exemple en reconnaissance d'images ou en traduction automatique. Ces avancées ont valu à leurs promoteurs, Yoshua Bengio, Georges Hinton et Yan Le Cun, l'attribution du prix Turing en 2019.

2.2 Algorithmes d'apprentissage

Compte tenu des besoins dans le domaine de la santé, **cet article est focalisé sur cette catégorie d'algorithmes dits d'apprentissage automatique (*machine learning, ML*)** qui représente les utilisations très majoritaires de l'IA du quotidien. Schématiquement, le ML se divise principalement en quatre classes d'algorithmes répondant à quatre objectifs :

- apprentissage ou classification *non supervisé*, lorsqu'aucun objectif à atteindre n'est a *priori* connu : reconnaissance de classes ou mesure quantitative. Il peut être question de débruiter ou déflouter une image, de rechercher des groupes homogènes (taxinomie, segmentation ou *clustering*) dans une population décrite par un ensemble de variables ou caractéristiques, comme par exemple segmenter en marketing des comportements pour la gestion de la relation client, définir des classes homogènes de patients au regard de leurs analyses biologiques ;

- apprentissage par *renforcement*, lorsque l'algorithme, disposant de règles de base, apprend en optimisant une fonction sous forme d'objectif ou récompense par des successions d'essais / erreurs au cours de la réception d'un flux de données ou d'expérimentations séquentielles. Tel est par exemple le cas de l'algorithme AlphaZero (Silver *et al.,* 2017) pour jouer au go ou aux échecs, ou encore celui de bandits manchots pour les systèmes de recommandation des sites de vente en ligne ;

- détection d'anomalie ou classification à une classe ou découverte de nouveautés ;

- apprentissage *supervisé* ou *statistique (statistical learning)* (James *et al.*, 2013), lorsqu'il est question de modéliser, expliquer et principalement prévoir la valeur d'une quantité ou celle d'une classe.

2.3 Apprentissage statistique ou IA empirique

C'est principalement ce dernier type d'apprentissage qui envahit notre quotidien, lorsqu'il s'agit d'attribution ou risque d'un crédit, d'analyse automatique de textes (CV ou *tweets*), d'évaluation du risque de récidive d'un accusé ou détenu, de la gestion des patrouilles de police en prévoyant les zones les plus probables de délits, d'aides au diagnostic médical… Les applications en sont innombrables, corrélatives à une production académique considérable.

De façon générale, un modèle est estimé ou un algorithme entraîné pour rendre visibles des relations entre une variable Y cible (le risque, le diagnostic…) et un ensemble de variables ou caractéristiques (*features*) dites explicatives X^j $(j=1,\ldots,p)$: caractéristiques socio-économiques, biologiques… Toutes ces variables (Y,X^j) sont mesurées, observées, sur un ensemble $i=1,\ldots,n$ d'individus ou *instances* appelé échantillon d'*apprentissage* ou d'entraînement. Une fois un modèle estimé ou un algorithme entraîné sur ces données, la connaissance d'un vecteur x_0, contenant les observations des variables X^j pour un nouvel individu, permet d'en déduire une prévision de la valeur ou de la classe y_0 le concernant. Le modèle ou l'algorithme calcule automatiquement cette valeur y_0 en combinant, en fonction de l'algorithme utilisé, celles y_i observées sur les individus présents dans la base d'apprentissage et proches de x_0, en un certain sens, au regard des valeurs x_i^j. Autrement dit, la prévision d'une nouvelle situation et donc la décision qui en découle, est construite automatiquement à partir des situations lui ressemblant le plus dans la base d'apprentissage et dont les décisions sont déjà connues. Le principe repose sur la stationnarité des données : la loi apprise sur l'échantillon d'apprentissage est la même que celle des données que l'on veut tester. En conséquence, l'apprentissage statistique n'*invente rien*, il reproduit un modèle connu et le généralise aux nouvelles données, *au mieux* selon un critère spécifique d'ordre statistique à optimiser. Plus on possède de données, meilleure sera la connaissance fournie par ce modèle. Ceci souligne le rôle fondamental joué par les données et donc le succès des grands acteurs d'internet et des réseaux sociaux qui bénéficient d'une situation de monopole sur des masses considérables de données comportementales des internautes pour les traduire en profilage et donc en recettes publicitaires. Transposé au domaine de la santé, où l'objectif est une prise en compte toujours plus fine de la complexité du vivant, le premier enjeu est l'accès à de grandes masses de données personnelles excessivement sensibles, objet de toutes les convoitises.

2.4 Statistique inférentielle versus apprentissage statistique

Deux objectifs doivent être clairement distingués dans les applications, tant de la statistique que de l'IA en santé pour lever une ambiguïté trop répandue.

Le premier objectif est celui *explicatif* de la statistique inférentielle, poursuivi par la mise en œuvre de tests, afin de montrer l'*influence d'un facteur* en contrôlant le *risque d'erreur*, soit le risque de rejeter à tort une hypothèse dite H_0 et donc de considérer que le facteur à un impact, alors qu'il n'en a pas. C'est le cas typique des essais cliniques de phase III, durant lesquels une molécule est prescrite en double aveugle à un groupe témoin, tandis que le groupe contrôle reçoit un *placebo*. Pour beaucoup de disciplines académiques, le test statistique constitue un outil de preuve scientifique même si son usage, parfois abusif, est mis en cause voire controversé à cause du manque de reproductibilité de trop nombreuses publications scientifiques (Ioannidis, 2016).

Le deuxième objectif est *prédictif*, en utilisant des modèles statistiques classiques ou les

algorithmes d'apprentissage automatique plus récents et sophistiqués. Deux sous-objectifs sont à considérer ; le premier est une prévision avec explication des résultats, de la façon dont les variables X^j influent sur la cible ou variable réponse Y. Le deuxième est une prévision brute sans recherche ou possibilité d'explication. Mais dans les deux cas, le *data scientist* sélectionne le modèle ou algorithme minimisant une estimation ou mesure d'une *erreur de prévision* qui contrôle le *risque d'erreur* de la décision qui en découle. *In fine* l'erreur de prévision de l'algorithme sélectionné est estimée sur un *échantillon test indépendant*, différent de l'*échantillon d'apprentissage* sur lequel il a été entraîné ; c'est aussi à la base de toute procédure de certification précédant sa mise en exploitation.

Il y a donc, selon les objectifs, deux types de risque ou d'erreur. Celui de se tromper en affirmant qu'un facteur est influent et celui de se tromper de décision à cause d'une erreur de prévision. Laissons la question, largement débattue par ailleurs (Ioannidis, 2016), de la pertinence des tests statistiques pour nous focaliser sur celle de la qualité de prévision plus spécifique à l'IA.

Il existe de très nombreux critères ou métriques pour évaluer une erreur de prévision. Ce peut être un simple *taux d'erreur* pour la prévision d'une variable binaire : tissus pathologiques ou sains, une *erreur quadratique moyenne* pour une variable Y quantitative. Dans beaucoup de publications du domaine de la santé, il est fait référence à l'aire sous la courbe ROC (*Area Under the Curve, AUC*) pour évaluer la qualité d'un algorithme pour une prévision binaire. Ce critère issu du traitement du signal nécessite quelques explications rappelées en annexe1.

2.5 Facteurs de qualité d'une prévision

Plus précisément, quels sont les composants d'un modèle statistique ou algorithme d'apprentissage qui sont déterminants pour la qualité de prévision et donc pour les risques d'erreur de la décision qui en découle ?

Le point fondamental pour la qualité ou robustesse, voire la certification d'un algorithme d'apprentissage statistique, est, en tout premier lieu, la *qualité des données* disponibles, ainsi que leur *représentativité* du domaine d'étude ou d'application concerné. Les données d'entraînement de l'algorithme sont-elles bien représentatives de l'ensemble des situations ou cas de figure susceptibles d'être, par la suite, rencontrés lors de l'exploitation de l'algorithme ? Il s'agit d'anticiper une capacité de généralisation de son usage. En effet, si des groupes ou des situations sont absents ou simplement sous-représentés, c'est-à-dire si les données sont, d'une façon ou d'une autre, *biaisées*, le modèle ou l'algorithme qui en découle ne fait que reproduire les biais ou s'avère incapable de produire des prévisions correctes de situations qu'il n'a pas suffisamment apprises lors de son entraînement. Ce problème est très bien référencé dans la littérature et souligné dans les rapports et guides éthiques. C'est même un vieux problème déjà formalisé en statistique pour la constitution d'un échantillon relativement à une population de référence en planification d'expérience ou en théorie des sondages. Ce n'est pas parce que les données sont volumineuses, déjà acquises, qu'il faut pour autant tout prendre en compte ou ne pas se préoccuper d'en acquérir d'autres. Considérons l'exemple typique de la prévision d'événements rares mais catastrophiques. Un algorithme naïf, pour ne pas dire trivial, conduit à un très faible taux d'erreur s'il ne prévoit aucune occurrence de l'événement rare, mais est inutile voire dangereux. L'expérience du *data scientist* le conduit alors à sur-représenter (sur-échantillonnage) les événements rares, ou sous-échantillonner ceux très fréquents ou encore à introduire des pondérations dans le choix de la fonction objectif à optimiser. Ces pondérations dépendent de l'asymétrie des coûts, à évaluer par des *experts métier*, d'un faux positif ou prévision à tort d'un événement exceptionnel, relativement au coût induit par un faux négatif qui n'anticipe pas la catastrophe.

La précédente question concerne la représentativité des individus ou situations présentes

dans la base d'entraînement relativement à une *population théorique de référence*. La deuxième soulève celle du choix ou de la disponibilité des caractéristiques ou variables observées sur ces individus. Elle peut se formuler de la façon suivante : les *causes effectives* de la cible ou variable Y à modéliser, ou les variables qui lui seraient très corrélées, sont-elles bien prises en compte dans les observations ? Dans le même ordre d'idée et avec les mêmes conséquences, des mesures peuvent être erronées, soumises à du bruit. Ces questions ne sont pas plus faciles à résoudre que celles de représentativité précédentes, car il n'est pas possible de pallier une absence d'information ou rectifier des erreurs de mesures ou de labellisations, mais il est plus simple d'en circonscrire les conséquences en estimant précisément les erreurs d'ajustement du modèle ou d'entraînement de l'algorithme puis celles de prévision ; elles resteront plus ou moins importantes mais évaluables, quel que soit le nombre de variables prises en compte ou le volume des données accumulées.

Plus précisément, la taille de l'échantillon ou le nombre d'instances de la base d'apprentissage intervient à deux niveaux sur la qualité de prévision. La taille nécessaire dépend, d'une part, de la complexité de l'algorithme, du nombre de paramètres ou de poids qui en définit la structure et, d'autre part, de la variance du bruit résiduel ou erreur de mesure. Un algorithme est entraîné, en moyenne, et la taille de l'échantillon doit être d'autant plus grande que la variance de l'erreur de mesure est importante. Les réseaux de neurones profonds appliqués à des images de plusieurs millions de pixels sont composés de dizaines de couches pouvant comporter des millions de paramètres ou poids à estimer ; ils nécessitent des bases de données considérables.

Attention, lorsque n est très grand (*big data*), le modèle peut être bien estimé car c'est une *estimation en moyenne* dont la précision s'améliore proportionnellement avec la racine de n. En revanche, une prévision individuelle est toujours impactée par le bruit résiduel du modèle, sa variance, quelle que soit la taille de l'échantillon. Aussi, même avec de très grands échantillons, la prudence est de mise quant à la précision de la prévision d'un *comportement individuel* surtout s'il est mal ou peu représenté dans la base : acte d'achat, acte violent, défaut de paiement, occurrence d'une pathologie.

En résumé, les applications quotidiennes de l'IA sont l'exploitation d'algorithmes d'apprentissage statistique, particulièrement sensibles à la qualité des données d'entraînement. Leur quantité est importante mais ne suffit pas à garantir la précision de prévisions individuelles qui doit être évaluée avec soin, afin de garantir, certifier, l'usage d'un algorithme. Malgré les abus de communication, l'IA ne se résume pas à l'utilisation de l'apprentissage profond. Le succès très médiatisé de certaines de ses applications ne doit pas laisser croire que ces relativement bons résultats en reconnaissance d'images ou traduction automatique sont transposables à tout type de problème.

Enfin, à l'exception des modèles statistiques élémentaires car linéaires ou à celle des arbres binaires de décision, les algorithmes d'apprentissage statistique sont opaques à une interprétation fine et directe de l'influence des caractéristiques d'entrée ou variables explicatives sur la prévision de la variable cible Y. Ce point soulève des problèmes délicats lorsqu'il s'agit de fournir l'*explication intelligible* d'une décision automatique. Des pistes de solutions ou d'aide à des solutions existent (Barredo Arrieta *et al.*, 2020) mais nous verrons ci-dessous que les applications de l'IA en santé appréhendent cette question de façon spécifique.

3. Cadre juridique de l'IA en Santé

Schématiquement, trois questions sont à prendre en considération pour préciser les frontières de l'action juridique :

- **Comment rendre compte de décisions** et en préciser les responsabilités, lorsqu'elles sont issues d'algorithmes souvent caractérisés par leur opacité ?

- **Quels sont les risques de discrimination** envers des personnes protégées ou groupes sensibles ?
- **Comment évaluer l'équilibre bénéfice / risque** entre l'intérêt public, d'une part, et le risque pour la vie privée des personnes touchées par l'utilisation de leurs données personnelles, d'autre part ?

3.1 Redevabilité et information versus opacité

L'article L.1111-4 du code de la santé publique précise que : « *Aucun acte médical ni aucun traitement ne peut être pratiqué sans le consentement libre et éclairé de la personne et ce consentement peut être retiré à tout moment* ». Le projet de loi bioéthique de 2019 intègre un article 11 spécifique sur l'utilisation de l'IA dans un cadre médical. Une fois voté, il devrait être intégré au chapitre I^{er} du titre préliminaire du livre préliminaire de la quatrième partie du code de la santé publique, et complété par un article L. 4001-3 ainsi rédigé :

« *I. – Lorsque pour des actes à visée préventive, diagnostique ou thérapeutique est utilisé un traitement algorithmique de données massives, le professionnel de santé qui communique les résultats de ces actes informe la personne de cette utilisation et des modalités d'action de ce traitement.*

II. – L'adaptation des paramètres d'un traitement mentionné au I pour des actions à visée préventive, diagnostique ou thérapeutique concernant une personne est réalisée avec l'intervention d'un professionnel de santé et peut être modifiée par celui-ci.

III. – La traçabilité des actions d'un traitement mentionné au I et des données ayant été utilisées par celui-ci est assurée et les informations qui en résultent sont accessibles aux professionnels de santé concernés. »

Seraient ainsi consacrés un droit à l'information sur l'utilisation d'un dispositif d'IA et un droit à une intervention humaine, celle du professionnel de santé, pour garantir le respect du droit à l'information sur les actes médicaux réalisés (à visée préventive, diagnostique ou thérapeutique) qui fonde le consentement libre et éclairé du patient.

En outre, le rôle du médecin n'est pas seulement d'informer le patient sur le recours à l'IA mais ce dernier doit aussi avoir la capacité d'intervenir sur l'utilisation du traitement algorithmique en modifiant les paramètres. Pour que le médecin puisse prendre des décisions en connaissance de cause, la traçabilité des actions est prévue. Cette interaction homme-machine pose naturellement la question de la responsabilité du médecin. Mais dès lors que le médecin reste au centre de la relation de confiance avec le patient, qu'il continue d'assumer une obligation d'information et qu'il reste maître des choix et décisions prises, la machine doit être considérée comme une simple aide à la décision qui ne remplace pas le médecin et ne modifie en rien les règles de responsabilité. En l'état actuel, les règles de responsabilité médicale applicables au médecin ne sont donc pas modifiées par le recours à un traitement algorithmique. En principe, le médecin assume une obligation de soin qui est une obligation de moyen, et non de résultat. Il n'engage sa responsabilité qu'en cas d'erreur fautive ayant entraîné un dommage.

Si l'obligation d'information concernant le recours à un dispositif d'IA ne cause pas de problème particulier, il sera sans doute plus difficile de garantir que le médecin pourra informer le patient des « modalités d'action de ce traitement » algorithmique. Encore faudra-t-il que le médecin le comprenne lui-même, ce qui pourra s'avérer difficile voire impossible dans certaines situations. Les conditions de déploiement des algorithmes doivent donc tenir compte de ces exigences et les entreprises privées qui proposeront leur système d'IA devront par conséquent expliquer voire former les médecins à la bonne utilisation de ces outils pour que ces derniers puissent à leur tour informer les patients, au moins sommairement, de la façon dont ils fonctionnent.

Au-delà de l'obligation d'information, il paraît de toute façon pertinent que les médecins puissent maîtriser un minimum ces outils pour que le médecin ait lui-même confiance et qu'ils deviennent de véritables aides à la décision.

Au demeurant, il paraît nécessaire d'interdire l'utilisation des méthodes algorithmiques opaques en matière de santé. Sur le modèle de ce que prévoit la loi s'agissant des décisions administratives automatiques, il pourrait être imposé que le responsable de traitement doive s'assurer de la maîtrise du traitement algorithmique et de ses évolutions, afin de pouvoir expliquer, en détail et sous une forme intelligible à la personne concernée, la manière dont le traitement a été mis en œuvre à son égard (art. L.311-3-1 du code des relations entre le public et l'administration). Ne peuvent être alors utilisés des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement (voir l'interprétation du Conseil constitutionnel dans sa décision n° 2018-765 DC du 12 juin 2018 (pt 71)). De telles dispositions ne supprimeraient pas mais réduiraient sensiblement les risques d'opacité.

3.2 Risque de discrimination

Les discriminations sont pénalement sanctionnées à l'article 225-1 al. 1er du code pénal qui définit la discrimination directe comme étant « *toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée* ». La loi énumère ainsi très largement les critères exhaustifs à prendre en compte pour rechercher si une discrimination directe a été commise. Une telle discrimination est intentionnelle et sera probablement plus facile à prouver.

L'alinéa 2 vise la discrimination indirecte qui est « *une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés* ». La discrimination indirecte est plus difficile à prouver car elle est non intentionnelle. Les traitements algorithmiques sont susceptibles d'être ainsi qualifiés car les discriminations peuvent être systémiques, par exemple en raison des données d'apprentissage biaisées utilisées pour entraîner le système d'IA, et avoir des répercussions sur des individus ou groupes d'individus. La preuve risque d'être difficile à rapporter, aussi est-il particulièrement fondamental en matière de santé d'être exigeant sur les conditions de constitution et utilisation de ces outils.

La notion de discrimination individuelle est reprise à l'article L.1110-3 du code de santé publique, selon lequel « *aucune personne ne peut faire l'objet de discriminations dans l'accès à la prévention ou aux soins* » (al. 1er). En outre, « *un professionnel de santé ne peut refuser de soigner une personne pour l'un des motifs visés au premier alinéa de l'article 225-1 du code pénal ou à l'article 225-1-1 du code pénal ou au motif qu'elle est bénéficiaire de la protection complémentaire en matière de santé prévue à l'article L.861-1 du code de la sécurité sociale, ou du droit à l'aide prévue à l'article L.251-1 du code de l'action sociale et des familles ». On peut alors s'interroger sur la place que prendront les dispositifs d'IA et l'impérieuse nécessité qu'il y aura à ce que les données utilisées ne soient pas biaisées, au risque de constituer une discrimination « dans l'accès à la prévention ou aux soins* ». Surtout, on peut se demander comment prouver la discrimination.

3.3 Bénéfice d'intérêt public versus Risque individuel

Alors que les données personnelles sont protégées dans l'Union européenne par le RGPD et en droit français par la loi informatique et libertés (LIL3), l'article L.1461-3 du code de santé publique organise un régime national d'accès aux données à caractère personnel du système national des données de santé (SNDS) pour permettre des traitements suivant une finalité mentionnée au III de l'article L.1461-1 et répondant à un **motif d'intérêt public**. La mise à disposition des données peut se faire pour contribuer : (i) à l'information sur la santé ainsi que sur l'offre de soins, la prise en charge médico-sociale et leur qualité ; (ii) à la définition, à la mise en œuvre et à l'évaluation des politiques de santé et de protection sociale ; (iii) à la connaissance des dépenses de santé, des dépenses d'assurance maladie et des dépenses médico-sociales ; (iv) à l'information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ; (v) à la surveillance, à la veille et à la sécurité sanitaires ; (vi) à la recherche, aux études, à l'évaluation et à l'innovation dans les domaines de la santé et de la prise en charge médico-sociale.

Le décret n° 2016-1871 du 26 décembre 2016 relatif au traitement de données à caractère personnel dénommé « système national des données de santé » fixe les règles de gouvernance et désigne les organismes autorisés à accéder de manière permanente aux données du SNDS, en fonction des missions de service public qu'ils remplissent. Tel est, entre autres, le cas de la Direction générale de la santé, des Agences régionales de santé, de l'Agence nationale de santé publique, de l'Agence nationale de sécurité du médicament et des produits de santé, l'Institut national du cancer, de l'INSERM, des équipes de recherche des CHU et des centres de lutte contre le cancer (CSP, art. R. 1461-12). Le décret définit l'étendue de cette autorisation par différents critères, tels que la profondeur historique, l'aire géographique, les caractéristiques d'une population, ainsi que la possibilité ou non d'utiliser dans un même traitement des identifiants potentiels qui permettraient d'accroître le risque de ré-identification.

Le décret organise aussi un accès aux données du SNDS soumis à autorisation de la CNIL à des fins de recherche, étude ou évaluation dans le domaine de la santé par les organismes non listés dans le décret (notamment organismes privés) et les organismes habilités à accéder de façon permanente au SNDS qui dépasseraient les limites fixées par le décret. Un autre décret n° 2016-1872 du 26 décembre 2016 précise les modalités de fonctionnement de l'Institut National des Données de Santé (INDS) et du Comité d'Expertise pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CEREES). Ce comité reprend une partie des missions du CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé) et se prononce sur la mise en œuvre de tout traitement de données à caractère personnel ayant pour finalité la recherche, l'étude ou l'évaluation dans le domaine de la santé et n'impliquant pas la personne humaine. L'INDS est en lien direct avec le CEREES, afin de fournir un avis à la CNIL sur la cohérence entre la finalité de l'étude proposée, la méthodologie présentée et le périmètre des données auxquelles il est demandé accès.

En résumé, la loi met en place un Système National des Données de Santé, devenu ensuite le *Health Data Hub*, qui peut, sous réserve d'assurer la confidentialité des données et donc des personnes concernées, donner l'accès aux données pour différents objectifs dont celui de recherches scientifiques présentant un intérêt public substantiel. En plus de devoir préciser les conditions de sécurité et confidentialité des données pour éviter les risques de ré-identification, la question qui en découle directement est de savoir ce qu'est un « **intérêt public substantiel** » de la recherche en santé. Une balance des intérêts doit ici se mettre en place entre la protection des données et l'intérêt de la recherche. Le standard juridique de l'« intérêt public substantiel » est une notion floue peu claire mais qui permet une souplesse de mise en œuvre par une interprétation au cas par cas des bénéfices et des risques.

Un raisonnement comparable est intégré au sein même de la Loi Informatique et Liberté n° 2018-493 du 20 juin 2018 s'agissant du traitement des données à caractère personnel dans le domaine de la santé. L'article 66.I. prévoit que de tels traitements « *ne peuvent être mis en œuvre qu'en considération de la finalité d'intérêt public qu'ils présentent. La garantie de normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux constitue une finalité d'intérêt public* ».

3.4 Quelle évolution du cadre réglementaire ?

Le livre blanc (Commission Européenne, 2020) annonce un changement de paradigme sur le fondement des lignes directrices (High Level Expert Group, 2019) qui s'achèvent par une liste d'évaluation (pilote) *ex ante* qui constituera le dossier obligatoire et indispensable à une expertise ou audit d'un système d'IA. Cette liste de questions couvre les 7 points éthiques fondamentaux identifiés : action humaine et contrôle humain, robustesse technique et sécurité, respect de la vie privée et gouvernance des données, transparence, non-discrimination et équité, bien-être sociétal et environnemental, utilité, responsabilité. Ce n'est pas le lieu de discuter la pertinence des 10 pages de questions auxquelles le responsable d'un système d'IA devra répondre mais bien celui de souligner le renversement de la charge de preuve. Alors qu'il serait très difficile pour ne pas dire impossible à un usager d'apporter la preuve qu'il est victime par exemple d'une discrimination algorithmique, ce sera au responsable du traitement de montrer qu'il a pris les mesures nécessaires afin d'éviter des biais sources de discrimination (obligation de moyens).

La Commission européenne n'a pas encore proposé un cadre légal, mais il est important de noter qu'en matière de santé, des organismes de certification en anticipent le principe. Outre aux États-Unis où la FDA (Health, 2019) a posé un cadre pour l'autorisation de commercialisation de systèmes d'IA d'aide au diagnostic, des réflexions sont aussi menées en France au travers du guide de la CNEDiMTS (commission nationale d'évaluation des dispositifs médicaux et des technologies de santé) (Haute Autorité de Santé, 2020) pour le dépôt d'un dossier de remboursement des DSC (dispositifs de santé connectés) embarquant de l'IA.

4. Domaines de santé concernés par L'IA

Le projet de loi bioéthique de 2011 ne fait pas mention d'IA mais rend nécessaire, pour sa révision périodique, la réunion d'États Généraux de la bioéthique qui ont produit un rapport (France, 2018). Ce rapport aborde neuf points dont six ont des implications sociétales fondamentales : la procréation assistée, la recherche sur l'embryon, les dons d'organes, la fin de vie, les neurosciences, l'environnement ; trois autres concernent indirectement ou directement les applications de l'IA : les bases de données de santé, la médecine génomique, la robotisation de la médecine. Cette section a pour but de préciser les quelques domaines de santé pour lesquels il semble le plus pertinent de s'intéresser aux impacts du déploiement de l'IA.

4.1 Bases de données

L'accès aux données est un préalable indispensable. La mise en place du Système National des Données de Santé (SNDS) (art. L.1461-3 du code de santé publique) est le résultat de la volonté politique d'ouvrir un *hub des données de santé* respectant par construction l'anonymat des patients. Il est composé du regroupement de la base SNIIRAM de l'assurance maladie, de celles des hôpitaux (PMSI), de la base INSERM des causes de décès, des données relatives au handicap et de celles détenues par les caisses d'assurance maladie complémentaire. L'accès à ces données est contrôlé par l'Institut National des Données de Santé (INDS) après avis de la CNIL.

Par ailleurs, d'autres sites régionaux se mettent en place pour regrouper les données hospitalières comme celui de la *clinique des données de santé* pilotée par le CHU de Nantes pour le grand ouest et qui fait appel à une société privée (*Wedata*[7]) pour la phase d'anonymisation. De plus, le Plan Investissement d'Avenir (PIA) (plan Médecine France Génomique 2025[8]) prévoit la mise en place de plateformes de séquençage à haut débit. Deux ont été sélectionnées à la suite de l'appel d'offre : SeqOIA (Paris) et AURAGEN (Lyon). Celles-ci ont pour mission de séquencer des dizaines de milliers de génomes chaque année.

Toutes ces bases et bien d'autres s'intègrent au projet national de *Health Data Hub*[9] (HDH) qui met en place une *pseudonymisation* des données : noms et adresses des patients sont supprimés et le code national d'inscription au registre des personnes physiques (NIRPP) est crypté par une fonction de hachage non réversible. Ce code devient une clef d'appariement des données de chaque patient pour la fusion des différentes bases mais ne permet pas de revenir au NIRPP initial.

Arrêt de la Cour de Justice de l'Union Européenne, projet de décret gouvernemental, avis de la CNIL, du Conseil d'État, opposition du Conseil de la Caisse nationale d'Assurance Maladie, le choix d'une société de droit américain (*Microsoft Azure*) pour l'hébergement du HDH soulève des problèmes même avec une localisation géographique française des données. Il est inutile de tenter d'intervenir dans ce débat à notre niveau mais il semble important de souligner que son ampleur obère d'autres questions qu'il serait dommageable de laisser dans l'ombre.

4.2 Médecine génomique

L'un des principaux battages médiatiques en santé concerne les médecines dites *translationnelles* et *4p* pour médecine *prédictive* d'un risque pathologique, *préventive* de ce risque, *participative* incluant le patient à la prévention et *personnalisée* ou de *précision* avec un traitement thérapeutique spécifique au patient. Cette précision ou personnalisation peut faire appel aux caractéristiques génétiques du patient et donc à la médecine dite *génomique*. La médecine *translationnelle* a pour objectif d'accélérer les applications de la recherche, donc des médecines précédentes, pour raccourcir le cycle de mise sur le marché d'un médicament. Elle nécessite de faciliter les échanges pluridisciplinaires ainsi qu'évidemment l'accès aux données médicales personnelles.

Schématiquement, deux types de bases de données génomiques sont constitués. Certaines, les plus récentes, enregistrent des séquences complètes de chaque génome ; 3,4 milliards de paires de base soit au minimum 3,5 GO par génome. Un génome complet comprend 1,5 % de parties codantes dans 26517 gènes protéiques. Les deuxièmes bases, de mises en place plus anciennes (Klein, 2005), se limitent à enregistrer pour chaque individu les présences/absences de variants génétiques ou mutations spécifiques appelées *single nucleotide polymorphism* (SNP). Jusqu'à 165 millions de SNP sont pris en compte pour chacun des milliers, ou millions d'individus de la base, auxquels sont associés un ensemble de phénotypes, c'est-à-dire la présence ou non de pathologies, des constantes biologiques. Ces bases permettent des études dites *pangénomiques* (*genomic wide association studies, GWAS*) en cherchant à mettre en relation variants génétiques ou mutations avec l'occurrence d'une pathologie.

Deux objectifs sont principalement poursuivis avec l'analyse de ces données. Le premier vise l'identification d'un élément potentiellement causal dans la survenue d'une maladie rare ou monogénique. Une mutation, éventuellement sur un gène, est associée à une fonction biologique défaillante et donc une pathologie. La mucoviscidose est un exemple type d'une telle

7. https://octopize-md.com/
8. https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/recherche-et-innovation/france-genomique
9. https://www.health-data-hub.fr/

maladie parmi plus de 8000 répertoriées dont beaucoup ne touchent que quelques familles dans le monde. Point important, la détection de la mutation responsable est obtenue par un test statistique qui détecte le facteur influent mais la prévision de la maladie concernée est validée par l'interprétation biologique ; elle n'est pas le résultat d'un algorithme d'apprentissage statistique. Un exemple spectaculaire d'une démarche de médecine personnalisée génomique translationnelle est fourni par le cas clinique (Kim *et al.*, 2019) d'une petite fille atteinte d'une maladie génétique dégénérative rare (Batten) et même exceptionnellement rare, unique dans le monde pour cette fillette, car la conséquence de deux mutations génétiques. Traitée à l'Hôpital de Boston, il a fallu un an pour déterminer et lui appliquer une thérapie génique qui n'est pas susceptible de la guérir mais au moins de réduire l'impact de la maladie dont le nombre de crises d'épilepsie par jour. Le coût global de cette démarche thérapeutique est resté confidentiel.

Un deuxième objectif vise à déterminer des facteurs génétiques de maladies multigéniques ou multifactorielles et souvent chroniques affectant une grande partie de la population. Cette démarche est basée sur des seuls éléments statistiques (tests) et pas sur l'analyse biologique des fonctions mises en cause car beaucoup trop de variants génétiques sont détectés. Elle occulte complètement les influences d'autres facteurs, environnementaux, épigénétiques, qui peuvent être largement prépondérants pour certaines pathologies. Ces insuffisances soulèvent de nombreuses critiques.

4.3 IA et robotisation de la médecine

Les États Généraux de la bioéthique font état des robots de microchirurgie, mais il n'en sera pas question ici. Nous allons nous focaliser sur d'autres types d'automatisation :

- aide au diagnostic par
- magerie médicale, électroencéphalogrammes (EEG), électrocardiogramme (ECG) et reconnaissance de formes par apprentissage profond ou deep learning,
- identification de biomarqueurs préventifs par études « omiques » ;
- aide aux choix thérapeutiques : e.g. IBM Watson ;
- surveillance des effets secondaires de médicaments à partir de la base SNIIRAM (Morel *et al*., 2019) ;
- suivi épidémiologique de grandes cohortes, telles que Constances10 (Zins *et al*., 2010).

Topol (2019) propose une revue assez exhaustive et enthousiaste des applications de l'IA en médecine mais nous nous limiterons aux quelques exemples illustrant les questions émergentes, juridiques ou éthiques. Ainsi, l'analyse automatique d'ECG relève des mêmes techniques d'apprentissage que l'analyse des images obtenues en radiologie ; seule cette dernière au développement viral est évoquée. La recherche de biomarqueurs transcriptomiques, protéomiques… d'une pathologie rejoint, d'un point de vue méthodologique, le débat sur la médecine génomique ; il n'est pas nécessaire de compléter. Suite à ce qui peut être considéré comme un coûteux échec (Ross and Swetlitz, 2018), *IBM* ne communique plus sur les applications de l'algorithme *Watson* en santé. C*et al*gorithme apprenait à partir de la littérature scientifique mais pas à partir de données personnelles sensibles. Son usage est principalement commercialisé dans le tertiaire, banque, assurance, c'est pourquoi nous le laisserons également de côté.

10. https://www.constances.fr/actualites/2019/js2019.php

5. Questions juridiques / éthiques de l'IA en santé

Comme évoqué précédemment, trois questions essentielles seront ici illustrées :

- biais et discrimination dans l'accès au soin ;
- consentement éclairé face à des algorithmes opaques ;
- balance bénéfice / risque entre intérêt de santé publique et ouverture des données.

5.1 Risques de discrimination des algorithmes d'apprentissage

La littérature académique propose (Žliobaitė, 2017) une très grande variété de critères ou définitions de la notion de biais et donc de discrimination des algorithmes d'apprentissage. Mais, comme le font remarquer Friedler *et al.* (2019), beaucoup sont très corrélés voire redondants et même pour certains incompatibles (Chouldechova, 2017). Nous nous limiterons à trois niveaux possibles, donc trois indicateurs de types de biais les plus régulièrement évoqués dans la littérature. Ils sont faciles à estimer par des intervalles de confiance afin d'en intégrer la précision (Besse *et al.*, 2018) si les données sont disponibles et fournissent un premier tableau synthétique suffisamment exhaustif des risques encourus de discrimination.
Le premier, nommé *demographic equality* dans la littérature, concerne la reproductibilité et aussi le risque d'amplification ou d'exacerbation de certains biais présents dans les données d'apprentissage. Comme dans beaucoup d'autres domaines, comme l'emploi, le crédit, le logement, les données thérapeutiques sont empreintes de biais de société. Lee *et al.* (2019) mettent ainsi en évidence dans une étude portant sur 85 millions de patients, que la gestion de la douleur par des antalgiques dépend de l'origine ethnique des patients. En toute logique, des algorithmes entraînés à partir de telles données reproduisent les biais voire les renforcent en se comportant donc de façon discriminatoire. Obermayer et Mullainathan (2019) dissèquent ainsi les biais ethniques produits par un algorithme qui guide les choix thérapeutiques de 70 millions de patients aux USA. Lorsque le système prévoit qu'un patient aura des besoins de soins futurs de santé particulièrement complexes et intensifs, il est inscrit à un programme qui fournit des ressources supplémentaires et une plus grande attention de la part de prestataires qualifiés ainsi qu'une aide à la coordination de ses soins. Les auteurs mettent en évidence un biais raciste en montrant comment les patients d'origine caucasienne, ayant le même état de santé que les patients d'origine afro-américaines, sont beaucoup plus susceptibles d'être inscrits dans le programme de gestion des soins et de bénéficier de ses ressources. Il s'agit là d'un *cas de prévision auto-réalisatrice* comme ceux dénoncés par O'Neil (2016). En synthétisant différents cas de sources de discrimination en santé, Vyas *et al.* (2020) ouvrent le difficile débat sur la pertinence ou non d'inclure l'origine ethnique dans les algorithmes cliniques.

Le deuxième indicateur est nommé *overall error equality*. C'est souvent la conséquence d'un autre type de biais initial lié à la mauvaise représentativité des données. Si un sous-groupe est sous-représenté, la prévision le concernant sera de moins bonne qualité. Ce biais est bien connu pour les applications de reconnaissance faciale (Buolamwini and Gebru, 2018). Particulièrement présent dans les données pangénomiques, ce biais fait que nous ne sommes pas tous égaux devant une médecine de précision qui personnaliserait les traitements à partir de considérations génétiques. En effet, la population d'ascendance blanche européenne (Popejoy and Fullerton, 2016) est présente à 96 % dans les bases génomiques en 2009 et encore à 81 % en 2016. Cette récente évolution est très majoritairement due au développement massif de campagnes de séquençage en Chine et donc sur des populations d'origine ethnique très spécifique. D'autres sources de problèmes sont aussi relevées dans ces données conduisant à d'autres risques de biais. Alors que beaucoup de pathologies (Pulit *et al.*, 2017) dépendent largement du sexe, cet aspect est négligé : les possibles mutations du chromosome X sont très rares (Chang *et al.*, 2014) dans ces bases et le chromosome Y en est absent. Enfin, la grande abondance de personnes relativement âgées et de leurs pathologies afférentes, ainsi que l'absence de prise en compte des facteurs environnementaux, biaisent l'étude des risques pathologiques des

patients jeunes.

Les algorithmes d'IA en santé ne semblent pas ou pas encore concernés, à première vue, par un troisième niveau de discrimination : *equality of odds*. Cet indicateur est à la base d'une vive controverse aux USA sur le caractère discriminatoire du logiciel *Compas* de prévision du risque de récidive. La société diffusant ce logiciel affirme qu'il ne discrimine pas au regard des deux précédents indicateurs tout en minimisant le rôle d'un taux d'erreur élevé de l'ordre de 30 à 40%. Larson et Angwin (2016), du site d'investigation *Propublica*[11], ont suivi une cohorte de près de 7000 détenus libérés dont ils connaissaient le score *Compas* de récidive à leur libération ainsi que l'occurrence ou non d'une récidive dans les deux ans la suivant. L'analyse de ces données porte, entre autres, sur les matrices de confusion (annexe 1) croisant le score de récidive : haut *vs.* bas avec la présence *vs.* absence de récidive. Ces matrices révèlent des asymétries inversées selon l'origine ethnique des personnes : celles d'origines afro-américaines présentent des taux de faux positifs (absence de récidive malgré un score élevé) plus importants que les personnes d'origine caucasienne. Cause d'un retard de libération et donc d'une plus forte désocialisation ; c'est encore un risque de prévision auto-réalisatrice.

Le diagnostic de ces problèmes en santé consiste à définir, détecter, les sources de biais dont les types sont maintenant bien identifiés. Leur résolution, ou au moins leur prise en compte, intervient à deux niveaux ; celui académique de déontologie scientifique et donc éthique, et celui réglementaire des organismes de certification pour une commercialisation et une exploitation publique à grande échelle.

Au niveau amont de la recherche académique, les études épidémiologiques de santé publique sont basées sur des cohortes dont la constitution est opérée avec une rigueur essentielle. Citons le cas de la cohorte *Constances* (Zins *et al.*, 2010) dont la mise en place sélective de volontaires sur une longue période a permis de réunir un échantillon de 200 000 personnes représentatif de la population nationale. Il en est de même pour les études basées sur un sous-ensemble de la base SNIIRAM (Schwarzinger *et al.*, 2018). La partie la plus délicate du travail n'est pas la modélisation bien balisée mais l'extraction des données pour constituer un échantillon représentatif conduisant à des résultats et des conclusions valides pour la population.

Détecter puis corriger les différentes sources de biais d'une base de données est de la responsabilité déontologique des chercheurs, essentielle à leur éthique. Cela concerne en premier chef également les relecteurs des revues scientifiques. En aval, c'est le rôle des organismes de certification. Le guide de la HAS (Haute Autorité de Santé, 2020) pour la rédaction du dossier de demande de remboursement inclut un questionnaire qui poursuit les mêmes objectifs que ceux de la liste d'évaluation du guide des experts de la CE. Le responsable du traitement est tenu de décrire toutes les dispositions prises pour s'assurer de la fiabilité, la robustesse, l'équité, la redevabilité du système d'IA concerné ; protocole analogue à celui mis en place aux USA par la *FDA* pour autoriser la commercialisation des AI/ML-SaMD (*Artificial Intelligence and Machine Learning Software as a Medical Device*) (Health, 2019).

5.2 Consentement éclairé versus Opacité des algorithmes

Le rapport Villani (Villani *et al.*, 2018) affirme que « *l'ouverture des boîtes noires de l'IA est un enjeu démocratique* » mais sans laisser entrevoir un embryon de piste pour une démarche qui peut prendre des formes multiples selon l'objectif visé et le contexte ou domaine d'application. Comme signalé en section 2.4, deux champs d'application sont déjà à considérer en fonction du type d'approche, explicative ou prédictive, mise en place.

11. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Dans de très nombreuses applications en santé, l'objectif est explicatif : trouver le gène dont la mutation est responsable d'une maladie rare ; identifier des biomarqueurs pour un diagnostic anticipé ou, en épidémiologie, les facteurs de risque d'une pathologie déterminée. Dans un cas comme de l'autre, l'objectif *explicatif* montrant l'influence d'un ou de quelques rares facteurs est prioritaire. Ce sont donc des modèles statistiques classiques qui sont utilisés pour ce type d'application : tests d'hypothèses pour les données génomiques (Lindström *et al.*, 2017), modèle de régression logistique ou modèle de Cox de durée de vie en épidémiologie. Il ne s'agit pas d'algorithmes d'apprentissage donc pas réellement d'intelligence artificielle mais comme le battage médiatique ne fait pas la différence et surtout que *l'enjeu juridique de l'accès aux données personnelles* est le même, ces modèles sont assimilés. Néanmoins, comme ils sont des modèles linéaires, pas des boîtes noires, ils permettent des interprétations détaillées pour atteindre l'objectif. La consultation de la valeur d'un coefficient, voire même simplement de son signe, suffit à la compréhension du modèle et à orienter l'explication biologique de l'influence d'un facteur.

La question de l'*explicabilité* est nettement plus critique lors du déploiement d'un apprentissage profond et même dès l'utilisation d'un réseau de neurones ou d'un algorithme d'apprentissage statistique sophistiqué. Il s'agit alors d'un modèle sous-jacent *non linéaire* d'un réseau potentiellement très complexe d'interactions entre les variables. L'effet d'une variable explicative sur la variable cible Y ne peut plus être explicité de façon détaillée. Ne nous trompons pas, certes l'algorithme est complexe mais pas plus que la réalité sous-jacente à l'analyse d'une image ou de celle du vivant incluant des effets simples, des interactions, des boucles de contre-réaction… Prenons l'exemple de l'imagerie médicale pour laquelle de très nombreuses expérimentations d'apprentissage profond ont été déployées : plus de 35000 références ont été identifiées par Liu *et al.* (2019a). Schématiquement, deux types d'explication sont à prendre en compte selon qu'elle s'adresse au chercheur ou ingénieur qui met en place l'algorithme ou à leurs usagers : le patient sur qui une image a été acquise en vue d'un diagnostic et le médecin qui le soigne. Le chercheur a besoin de comprendre finement les modes opératoires de l'algorithme, afin d'en détecter les failles et y remédier : dans quelles circonstances et à la suite de quel défaut la prévision est-elle erronée ? Quelles sont les situations insuffisamment présentes dans la base d'apprentissage pour être correctement identifiées ? Le patient, comme son médecin, ne s'intéressent pas à ce niveau de détail mais évidemment à leurs conséquences sur la qualité de la prévision. Annoncer un diagnostic avec un taux d'erreur de 1 % ou de 30 % change tout pour le patient comme pour l'équipe médicale qui doit en déduire une stratégie thérapeutique et en expliquer les conséquences. Comme l'exprime également London (2019), l'explication au patient doit être focalisée sur le risque d'erreur de cette aide au diagnostic, de la même façon que chirurgien et anesthésiste doivent expliquer à leur patient les risques afférents à une opération afin que le patient puisse exprimer son choix de façon *libre* et suffisamment *éclairée*.

Cette communication du risque d'erreur nécessite une estimation précise et sans biais mais cette tâche est, en principe, inhérente à l'entraînement d'un algorithme ; elle ne peut être évacuée et fait logiquement partie du processus de certification imposé par la FDA (Health, 2019) ou la demande de remboursement de la HAS (Haute Autorité de Santé, 2020). En revanche, au plan académique, où seule l'éthique ou la déontologie scientifique est opérante, de trop nombreuses défaillances révèlent de sérieux manques de rigueur sous la pression de publication. Liu *et al.* (2019a) témoignent que très peu de publications respectent un protocole rigoureux basé sur des estimations des erreurs utilisant des échantillons indépendants de celui d'apprentissage ; c'est une insuffisance méthodologique, donc déontologique, sans conséquence thérapeutique directe mais financière car, mêmes publiées, des études ne s'avéreront pas reproductibles lors de nouvelles analyses tout aussi coûteuses. Cela conduit Liu *et al.* (2019b) à proposer un guide sur les précautions à prendre en lisant un article d'application de l'apprentissage automatique en imagerie médicale.

5.3 Intérêt public versus risques sur la confidentialité

Le dernier point est le plus complexe à analyser. Il nécessite d'évaluer les bénéfices attendus de la recherche scientifique en santé au regard des risques encourus par l'ouverture de l'accès à des données personnelles. La réglementation européenne prévoit la protection de la confidentialité de ces données mais la garantie totale peut, soit être difficile à assurer, soit conduire à une dégradation des données et donc de la qualité des résultats escomptés.

Le *premier volet* concerne les risques de ré-identification en fonction du procédé d'anonymisation mis en œuvre. L'article L.1461-4 du code de la santé publique dispose que les données ne doivent pas contenir le nom, l'identifiant NIRPP et l'adresse des personnes. Dans le cas de pseudonymisation du HDH, les NIRPP sont cryptés pour servir de clef d'appariement. Ces précautions sont largement insuffisantes pour anonymiser des données. Une ré-identification partielle, c'est-à-dire celle d'un sous-ensemble des personnes de la base de données, peut être obtenue à partir des seules informations précisant la date de naissance, le code postal et le sexe de ces personnes. Si en plus le nombre d'enfants est connu, l'unicité du profil et donc l'identification devient très probable. De nombreux auteurs analysent ces risques (Rubinstein and Hartzog, 2016) ou en font la démonstration (Rocher *et al.*, 2019 ; Narayanan and Shmatikov, 2008). Supprimer quelques informations est donc insuffisant, il est nécessaire d'apporter suffisamment de flou dans les données les plus personnelles, par exemple en discrétisant l'âge en tranches, la localisation en grandes zones, afin de contrôler le risque d'unicité dans la population et donc de ré-identification. D'autres stratégies comme la *confidentialité différentielle* (Dwork and Roth, 2013) consistent à simuler une part de données synthétiques en respectant les principales propriétés. Dans tous les cas, il s'agit de chercher un meilleur compromis entre risque de ré-identification et dégradation des analyses statistiques. Le CHU de Nantes propose de remplacer les données réelles par des données synthétiques simulées à partir de k plus proches voisins. Les simulations sont supposées suffisamment réalistes pour que les principales propriétés statistiques – distributions et corrélations des variables – et donc les principales qualités des analyses futures soient conservées, tout en rendant en principe impossible le retour aux données initiales et la ré-identification. La procédure semble intéressante mais, mise en œuvre par une entreprise privée, le descriptif détaillé, protégé par le secret commercial, n'est pas accessible. Plutôt que de proposer du code libre d'accès évaluable par un audit indépendant, il est regrettable d'ajouter une couche d'opacité limitant la confiance envers les données générées.

Par ailleurs, il est à noter que le floutage des données ou la construction de données synthétiques n'est pas applicable aux bases de données pangénomiques. Altérer les présences / absences de mutations rendraient ces données inutilisables car la proximité sur l'arbre phylogénétique (cousinage) permet au FBI de résoudre des affaires classées[12] mais n'induit pas des proximités au sens des pathologies concernées comme permettent de l'inférer des proximités au sens de mesures biologiques quantitatives. De plus, la connaissance d'une liste de SNP relativement restreinte d'une personne peut jouer le rôle d'une *empreinte génétique* (Robinson and Glusman, 2018) et constituer une clef d'accès unique à une base génomique même anonymisée et contenant des informations sensibles sur les pathologies de cette personne. C'est en France le rôle de la CNIL de s'assurer de la confidentialité des données, qu'elles soient anonymisées ou seulement pseudonymisées. Dans ce derniers cas (HDH), l'accès aux données doit être particulièrement restreint et protégé afin d'éviter toute fuite par négligence ou malveillance, fuite des dossiers médicaux mais aussi fuite discrète d'une information synthétique ou score personnel de niveau de santé susceptible d'être commercialisé auprès d'une banque, assurance…

12. https://www.oregonlive.com/crime/2019/01/portland-police-tie-texas-serial-killer-rapist-to-40-year-old-homicide-case-using-public-genealogy-data.html

Le *deuxième volet* en balance est celui de l'intérêt public ou bien commun au Canada, conséquence des recherches. L'intérêt d'une recherche académique est généralement évalué par le nombre et l'impact des publications qui en découlent. Néanmoins, cette évaluation bibliométrique impacte l'intérêt des chercheurs – promotion ou accès à des subventions – pas directement celui du public. Il s'agit d'évaluer des intérêts publics concrets et substantiels ; telle est la mission en France du Comité d'Experts de l'Intérêt Public (CEIP) de l'Institut National des Données de Santé (INDS). Ce dernier accorde l'accès à des projets de recherche spécifiques, après avis consultatif de la CNIL sur le volet de la confidentialité.

Parmi les exemples typiques d'application de l'IA en santé, quels sont ceux conduisant à des intérêts substantiels ou non ? Il serait bien trop long de dérouler une étude exhaustive du problème et seulement cinq cas illustratifs seront considérés. Les deux premiers sont le résultat de tests et modèles statistiques avec un objectif explicatif et non prédictif. Il ne s'agit pas formellement d'IA mais le point important à considérer est bien l'ouverture de l'accès aux données et aussi donc la pertinence des résultats obtenus, application d'un algorithme vedette d'IA ou pas.

Le *premier cas* concerne les études épidémiologiques classiques, maintenant appliquées à de très grandes cohortes, en utilisant des modèles statistiques explicatifs. L'analyse de la cohorte *Constances* conduit ainsi à des résultats substantiels présentés chaque année lors d'une journée scientifique[13]. La sécurité des données est essentielle mais il s'agit d'une pratique ancienne et reconnue de la recherche médicale qui ne fait que se déployer en considérant des cohortes d'effectifs nettement plus importants afin de pouvoir détecter (puissance des tests) des facteurs ou combinaisons complexes de facteurs aux effets moins prononcés.

Dans le *deuxième cas*, des batteries de tests statistiques sont appliquées sur les bases pangénomiques pour mettre en évidence la mutation du gène, ou de son promoteur, responsable d'une maladie rare. Pujol (2019), président de la Société Française de Médecine Prédictive et Personnalisée (SFMPP), en décrit les enjeux et intérêts substantiels. Il explicite le difficile débat sur l'opportunité des tests génétiques, très encadrés en France, et sur la pertinence des informations à communiquer aux couples aux différentes étapes de la conception d'un enfant. Cette réflexion est basée sur deux concepts :

- l'estimation statistique de la pénétrance d'une mutation ou probabilité de développer la maladie qui lui est associée : elle est ainsi de 100 % pour la mucoviscidose mais de 75 % de développer un cancer du sein pour une mutation d'un des gènes BRCA ;
- l'actionnabilité ou possibilités thérapeutiques médicales ouvertes par un diagnostic de risque associé à une mutation.

L'utilisation des bases génomiques à cette fin n'est pas remise en cause et est à l'origine du plan France Génomique 2025 incluant une sécurité des données également essentielle.

Le *troisième cas* concerne l'utilisation emblématique de l'apprentissage profond en imagerie médicale en vue d'automatiser le diagnostic ou plutôt l'aide à ce diagnostic. De très nombreuses publications, largement médiatisées, témoignent de leur efficacité : Esteva *et al.* (2017), De Fauw *et al.* (2018), Haenssle *et al.* (2018), Yala *et al.* (2019)... Une synthèse de ces très nombreux travaux (Liu *et al*., 2019a) alerte sur le manque de rigueur de beaucoup de comparaisons entre diagnostic automatique et humain ; celles validées par une évaluation rigoureuse de l'erreur sur des échantillons test indépendants permettent de conclure à une capacité de diagnostic comparable entre l'algorithme et un panel de spécialiste. La FDA (Health, 2019) propose un protocole de certification élaboré qui a permis d'autoriser la pré-commercialisation (Topol, 2019) de nombreux AI/ML-SaMD (*Artificial Intelligence and Machine Learning in a Software as a Medical Device*). Attention, ces dispositifs ne sont pas infaillibles. Comme cela est expliqué plus

13. https://www.constances.fr/actualites/2019/js2019.php

haut, un algorithme d'apprentissage même profond ne peut prévoir que ce qu'il connaît et a déjà rencontré. Ainsi, Oakden-Rayner *et al.* (2019) révèlent le cas d'un cancer du poumon très rare non détecté par une analyse d'image automatique alors qu'il s'agit d'un cas mortel. C'est typiquement ce qui rend indispensable, comme le prévoit la réglementation de la FDA, la mise en place d'une surveillance constante et rétroactive des dispositifs de santé, afin d'en compléter, si nécessaire, l'apprentissage. Dans un livre blanc[14] sur « *le monde des data, des algorithmes et de l'IA* », le Conseil National de l'Ordre des Médecins appelle à juste titre à une *éthique de la vigilance*.

Le *quatrième cas* est la recherche de protéines biomarqueurs, illustré par les résultats encourageants de Williams *et al.* (2019). Ils considèrent une cohorte de 17000 patients pour lesquels 5000 protéines plasmatiques sont dosées à l'aide d'une technologie récente (aptamères) plutôt que par spectrométrie (LC MS/MS). Sur ces données, des algorithmes d'apprentissage – régression avec pénalisation Lasso et *ridge*, machine à vecteurs supports, forêts aléatoires… – conduisent à des bonnes prévisions de l'état de santé du patient et des principaux risques cardiovasculaires, diabètes, meilleures que celles des modèles cliniques usuels. Néanmoins, ces résultats nécessiteraient d'être confirmés sur un autre jeu de données car la présélection des protéines et l'algorithme d'apprentissage ont été exécutés sur le même jeu de données au risque d'un biais de sélection déjà souligné par Ambroise et McLachlan (2002) sur des études transcriptomiques.

Le *cinquième cas*, la recherche sur les maladies multifactorielles utilisant des données pangénomiques, est nettement plus controversé quant à l'intérêt public qu'elle peut apporter alors que c'est celle qui, à terme, brassera le plus grand volume de données. Elle attire de plus la convoitise des acteurs majeurs du numérique qui ont tous des projets plus ou moins avancés dans ce secteur. Ces réserves sont exprimées dès 2010 dans les conclusions d'un texte[15] issu d'une réflexion de la Société Française de Génétique Humaine (Bernheim *et al.*, 2010) et cosigné par l'ensemble des sociétés savantes et associations professionnelles de génétique et génétique humaine :

> « *Si les études pangénomiques apportent une contribution essentielle à la connaissance scientifique, l'utilisation exclusive de l'information qui en résulte est dénuée de sens en matière de prédiction de santé. Elle conduit à une perception erronée du risque encouru par les individus. Il est du devoir de la communauté scientifique de ne pas servir d'alibi en matière de prédictions individuelles de risque pour les maladies multifactorielles à partir de la seule information génomique.* »

Ceci n'a pas pour autant bloqué les programmes de recherche avec la mise en œuvre de méthodes *et al*gorithmes plus sophistiqués. L'annexe 2 en propose une rapide revue montrant des résultats peu probants ou obtenus à la suite d'une démarche manquant de rigueur. Comme déjà évoqué en 2010, ces résultats ne peuvent servir de preuve d'un intérêt public substantiel ou d'alibi pour accéder à de grandes bases de données.

À ce manque de résultats, il faut ajouter dans l'autre plateau de la balance des risques accrus de ré-identification déjà mentionnés. Les données génomiques intègrent implicitement une empreinte génétique définissable à partir d'une sélection de SNP (Robinson and Glusman, 2018). Sans garantie drastique de sécurité, ces empreintes sont autant de clefs d'identification exploitables par des sociétés telles que *23andme*[16] ou les entreprises à qui les données sont

14. https://www.conseil-national.medecin.fr/sites/default/files/cnomdata_algorithmes_ia_0.pdf
15. http://atlasgeneticsoncology.org/Associations/Predictions_risques_maladies_multifactorielles.pdf
16. https://www.23andme.com/en-int/

vendues[17]. Notons qu'aux USA, sans les contraintes européennes légales du RGPD, les principaux acteurs montent des partenariats pour constituer de gigantesques bases de données de santé : *Verily Life Science* filiale d'*Alphabet* et *GSK*[18], *Aetion* et *Sanofi*[19], projet *Nithingale* de *Google* et *Ascension*[20]... Dernière étape, *Verily* signe un partenariat[21] avec *Swiss Re Corporate Solution* (société d'assurances) avec l'opportunité de développer des contrats individualisés à l'encontre du principe, basique en assurance, de mutualisation du risque.

6. Conclusion

La réglementation européenne et les lois nationales sont claires : la *discrimination* est interdite, le *consentement libre et éclairé* des patients doit être requis, sauf pour l'accès à des données personnelles lorsque l'*intérêt public* ou bien commun de la recherche est avéré. En Europe, les questions soulevées par l'utilisation d'algorithmes d'IA en santé sont donc en premier lieu moins d'ordre éthique que juridique et réglementaire.

Les disparités socio-économiques, géographiques et maintenant numériques dans l'accès aux soins sont connues. Le risque, bien identifié, est que des algorithmes de décision d'apprentissage proposant des aides automatiques à la décision viennent renforcer ces biais, en ajoutent d'autres et donc discriminent. La FDA (Health, 2019) comme la HAS imposent des processus adéquats de détection en continu de ces biais tout au long de l'utilisation des seuls dispositifs soumis à leur certification. Ce processus intègre une évaluation de leur risque d'erreur et donc des risques de mauvais diagnostic quand celui-ci est le résultat de l'exécution d'un algorithme opaque. L'information due à l'équipe médicale et au patient pour solliciter son consentement est avant tout de faire connaître l'origine de la décision, l'évaluation du risque d'erreur associée, ainsi que l'opportunité d'un diagnostic complémentaire.

En amont de l'exploitation de ces algorithmes et dispositifs de santé, la recherche doit être *scientifique* et donc ses résultats *reproductibles*. Un effort important doit être consenti par les acteurs de la recherche pour acquérir les compétences indispensables au déploiement d'algorithmes sophistiqués, puissants mais tellement sensibles à la qualité des données, leur représentativité. Il s'agit d'en maîtriser les limites tout en résistant à la pression de publication. Les données étudiées pouvant être confidentielles, il importe de rendre accessible les codes de calcul commentés (Donoho, 2017) afin de permettre une évaluation transparente de la démarche ; les outils actuels (*jupyter notebook*, dépôts *git*) le rendent facile. Le secret commercial n'est pas opposable car les données de l'apprentissage sont protégées et son résultat, le modèle, peut rester confidentiel.

Prenant en considération toute la complexité combinatoire du vivant, associant diversité des variants génétiques et diversité environnementale des conditions de vie, les algorithmes d'apprentissage statistique, même et surtout les plus sophistiqués, partent avec un lourd handicap pour atteindre les objectifs ambitieux d'une médecine personnalisée et prédictive des maladies chroniques multifactorielles. Évaluer globalement les facteurs de risque environnementaux ou génétiques d'une pathologie multifactorielle pour une population est une chose, prévoir très tôt, pour un individu, le risque qu'il déclenche une telle maladie, ou améliorer la prévision obtenue à partir de ses seuls paramètres cliniques en utilisant ses caractéristiques génomiques en est une autre. *Il y a une forme d'antinomie entre les principes de l'apprentissage statistique, basés sur des données, et les objectifs de la médecine prédictive personnalisée* ; au regard des

17. https://www.usinenouvelle.com/article/23andme-vend-l-integralite-des-donnees-genetiques-de-ses-clients-au-laboratoire-gsk-et-cree-la-polemique.N729654
18. https://www.usine-digitale.fr/article/apres-novartis-et-sanofi-verily-life-science-google-se-rapproche-de-gsk.N421917
19. https://www.clinicaltrialsarena.com/news/sanofi-and-aetion-to-integrate-real-world-data-platforms/
20. https://www.theguardian.com/technology/2019/nov/14/google-healthcare-data-ascension
21. https://www.bloomberg.com/news/articles/2020-08-25/alphabet-s-verily-plans-to-use-big-data-as-health-insurance-tool

caractéristiques génétiques et plus encore en croisant celles génétiques et environnementales, chaque humain est unique donc difficilement prédictible ; Keyes *et al.* (2015) met en évidence ces mêmes problèmes à l'aide de simulations.

L'ouverture et l'accès aux données de santé notamment génomiques pour la recherche doivent être conditionnés à des pratiques déontologiques très strictes : exiger des mesures draconiennes de sécurité dans la gestion de bases de données très sensibles, afin d'éviter toute faille de sécurité par incompétence ou même malveillance, exiger une démarche scientifiquement rigoureuse garante de la production de *résultats reproductibles* : identification et correction des biais de tout ordre, afin de produire des prévisions représentatives d'une population de référence à définir, contrôle des prétraitements pour ne pas rajouter de biais qui conduiraient à des situations irréalistes, estimation des erreurs de prévision (AUC) sur des échantillons réellement indépendants et représentatifs de l'usage projeté, publication des codes de calcul afin de faciliter les vérifications.

Ces quelques réflexions nous amènent à suggérer des recommandations à trois niveaux : accès aux données, déontologie de la recherche, réglementation des dispositifs de santé connectés. Ces recommandations sont regroupées dans le *tableau 1* inclus en introduction.

Références

Alaa A. M., T. Bolton, E. Di Angelantonio *et al.* (2019), « Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants », *PLoS ONE*, vol. 14, p. e0213653, https://doi.org/10.1371/journal.pone.0213653.

Ambroise C., G. J. McLachlan (2002), « Selection bias in gene extraction on the basis of microarray gene-expression data », *Proceedings of the National Academy of Sciences*, vol. 99, pp. 6562–6566, https://doi.org/10.1073/pnas.102102699.

Barredo Arrieta A., N. Díaz-Rodríguez, J. Del Ser *et al.* (2020), « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Information Fusion*, vol. 58, pp. 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

Bernheim A., C. Bourgain, A. Cambon-Thomsen *et al.* (2010), « Quelle valeur accorder aux prédictions de risques pour les maladies multifactorielles ? », Texte émanant de la Société Française de Génétique Humaine.

Besse P., E. del Barrio, P. Gordaliza, and J.-M. Loubes (2018), « Confidence Intervals for Testing Disparate Impact in Fair Learning », arXiv:180706362 [cs, math, stat].

Buchanan B. G. and E. H. Shortliffe (eds.) (1984), *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Mass.

Buolamwini J. and T. Gebru (2018), « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification », *in Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Chang D., F. Gao, A. Slavney *et al.* (2014), « Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases », *PLoS ONE*, vol. 9, p. e113684, https://doi.org/10.1371/journal.pone.0113684.

Chouldechova A. (2017), « Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments », *Big Data*, vol. 5, pp. 153–163, https://doi.org/10.1089/big.2016.0047.

Commission Européenne (2018), « Le règlement général sur la protection des données - RGPD ».

Commission Européenne (2020), commission-white-paper-artificial-intelligence-feb2020_fr.pdf, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf (Accessed 12 Feb 2021).

Darlington K. W. (2011), « Designing for Explanation in Health Care Applications of Expert Systems », *SAGE Open*, vol. 1, p. 21582440114086, https://doi.org/10.1177/2158244011408618.

De Fauw J., J. R. Ledsam, B. Romera-Paredes *et al.* (2018), « Clinically applicable deep learning for diagnosis and referral in retinal disease », *Nat. Med.*, vol. 24, pp. 1342–1350, https://doi.org/10.1038/s41591-018-0107-6.

Do C. B., D. A. Hinds, U. Francke, and N. Eriksson (2012), « Comparison of Family History and SNPs for Predicting Risk of Complex Disease », *PLoS Genet*, vol. 8, p. e1002973, https://doi.org/10.1371/journal.pgen.1002973.

Donoho D. (2017), « 50 Years of Data Science », *Journal of Computational and Graphical Statistics*, vol. 26, pp. 745–766, https://doi.org/10.1080/10618600.2017.1384734.

Dwork C. and A. Roth (2013), « The Algorithmic Foundations of Differential Privacy », *FNT in Theoretical Computer Science*, vol. 9, pp. 211–407, https://doi.org/10.1561/0400000042.

Esteva A., B. Kuprel, R. A Novoa *et al.* (2017), « Dermatologist-level classification of skin cancer with deep neural networks », *Nature*, vol. 542, pp. 115–118, https://doi.org/10.1038/nature21056

Eurostat (2017), « Code de bonne pratique de la Statistique européenne ».

Fjeld J., H. Hilligoss, N. Achten N *et al.* (2019), « Principled Artificial Intelligence », https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf (Accessed 11 Oct 2019).

France Comité Consultatif National d'Éthique (2018), « États Généraux de la Bioéthique : Rapport de Synthèse du Comité Consultatif National d'Éthique – Opinion du comité citoyen », La Documentation Française.

Friedler S. A., C. Scheidegger, S. Venkatasubramanian *et al.* (2019), « A comparative study of fairness-enhancing interventions in machine learning », *in Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19*, Atlanta, GA, USA, ACM Press, pp. 329–338.

Gim J., W. Kim, S. H. Kwak *et al.* (2017), « Improving Disease Prediction by Incorporating Family Disease History in Risk Prediction Models with Large-Scale Genetic Data », Genetics, vol. 207, pp. 1147–1155, https://doi.org/10.1534/genetics.117.300283.

Guichard C. (2018), « Affaire Cambridge Analytica : Facebook chute de près de 7 % en Bourse », *in Courrier international*, https://www.courrierinternational.com/article/affaire-cambridge-analytica-facebook-chute-de-pres-de-7-en-bourse (Accessed 9 Nov 2019).

Haenssle H.A., C. Fink, R. Schneiderbauer *et al.* (2018), « Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists », *Annals of Oncology*, vol. 29, pp. 1836–1842, https://doi.org/10.1093/annonc/mdy166.

Haute Autorité de Santé (2020), « Guide : LPPR Dépôt d'un dossier auprès de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé ».

Health C for D and R (2019), « Artificial Intelligence and Machine Learning in Software as a Medical Device », FDA.

High Level Expert Group (2019), « Ethics guidelines for trustworthy AI », *in Digital Single Market – European Commission*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Accessed 9 Nov 2019).

Ho D. S. W., W. Schierding, M. Wake *et al.* (2019), « Machine Learning SNP Based Prediction for Precision Medicine », *Frontiers in Genetics*, vol. 10, p. 267, https://doi.org/10.3389/fgene.2019.00267

Ioannidis J. P. A. (2016), « Why Most Clinical Research Is Not Useful », *PLoS Med*, vol. 13, p. e1002049, https://doi.org/10.1371/journal.pmed.1002049.

James G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, New York, NY, Springer.

Jobin A., M. Ienca, and E. Vayena (2019), « The global landscape of AI ethics guidelines », *Nature Machine Intelligence*, vol. 1, pp. 389–399, https://doi.org/10.1038/s42256-019-0088-2.

Kahn H. S. (2009), « Two Risk-Scoring Systems for Predicting Incident Diabetes Mellitus in U.S. Adults Age 45 to 64 Years », *Annals of Internal Medicine*, vol. 150, p. 741, https://doi.org/10.7326/0003-4819-150-11-200906020-00002.

Keyes K. M., G. Davey Smith, K. C. Koenen, and S. Galea (2015), « The mathematical limits of genetic prediction for complex chronic disease », *Journal of Epidemiology and Community Health*, vol. 69, pp. 574–579, https://doi.org/10.1136/jech-2014-204983.

Kim J., C. Hu, C. Moufawad El Achkar *et al.* (2019), « Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease », *The New England Journal of Medicine*, vol. 381, pp. 1644-1652, https://doi.org/10.1056/NEJMoa1813279.

Klein R. J. (2005), « Complement Factor H Polymorphism in Age-Related Macular Degeneration », *Science*, vol. 308, pp. 385–389, https://doi.org/10.1126/science.1109557.

Kraege V., J. Fabecic, P. M. Vidal *et al.* (2020), « Validation of seven type 2 diabetes mellitus risk scores in a population-based cohort. The CoLaus Study », *The Journal of Clinical Endocrinology & Metabolism*, vol. 105, n° 3, https://doi.org/10.1210/clinem/dgz220.

Larson J. and J. Angwin (2016), « How We Analyzed the COMPAS Recidivism Algorithm », *in ProPublica*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Accessed 10 Nov 2019).

Lee P., M. Le Saux, R. Siegel *et al*. (2019), « Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review », *The American Journal of Emergency Medicine*, vol. 37, pp. 1770–1777, https://doi.org/10.1016/j.ajem.2019.06.014.

Lindström S., S. Loomis, C. Turman *et al*. (2017), « A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts », *PLoS ONE*, vol. 12, p. e0173997, https://doi.org/10.1371/journal.pone.0173997.

Liu X., L. Faes, A. U. Kale *et al*. (2019a), « A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis », *The Lancet Digital Health*, vol. 1, pp. e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2.

Liu Y., P.-H. C. Chen, J. Krause, and L. Peng (2019b), « How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature », *JAMA*, vol. 322,p. 1806, https://doi.org/10.1001/jama.2019.16489.

London A. J. (2019), « Artificial Intelligence and Black-Box Medical Decisions: Accuracy *versus* Explainability », *Hastings Center Report*, vol. 49, pp. 15–21, https://doi.org/10.1002/hast.973.

Lopez B., F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real (2018), « Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction », *Artificial Intelligence in Medicine*, vol. 85, pp. 43–49, https://doi.org/10.1016/j.artmed.2017.09.005.

Mieth B., M. Kloft, J. A. Rodríguez *et al*. (2016), « Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies », *Scientific Reports*, vol. 6, pp. 1–14, https://doi.org/10.1038/srep36671.

Montanez C. A. C., P. Fergus, A. C. Montanez *et al*. (2018), « Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs », *in* 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, Rio de Janeiro, pp. 1–8.

Morel M., E. Bacry, S. Gaïffas *et al*. (2019), « ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection », *Biostatistics*, vol. 21, pp. 758–774, https://doi.org/10.1093/biostatistics/kxz003.

Narayanan A. and V. Shmatikov (2008), « Robust De-anonymization of Large Sparse Datasets », *in* 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, Oakland, CA, USA, pp. 111–125.

Oakden-Rayner L., J. Dunnmon, G. Carneiro, and C. Ré (2019), « Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging », arXiv:190912475 [cs, stat].

Obermeyer Z. and S. Mullainathan (2019), « Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People », *in Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19*, Atlanta, GA, USA, ACM Press, pp. 89–89.

O'Neil C. (2016), *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition), New York, Crown.

Patron J., A. Serra-Cayuela, B. Han *et al*. (2019), « Assessing the performance of genome-wide association studies for predicting disease risk », *PLoS ONE*, vol. 14, p. e0220215, https://doi.org/10.1371/journal.pone.0220215.

Popejoy A. B. and S. M. Fullerton (2016), « Genomics is failing on diversity », *Nature*, vol. 538, pp. 161–164, https://doi.org/10.1038/538161a.

Pujol P. (2019), *Voulez-vous savoir ? Ce que nos gènes disent de notre santé*, Paris, Editions humensciences.

Pulit S. L., T. Karaderi, and C. M. Lindgren (2017), « Sexual dimorphisms in genetic loci linked to body fat distribution », *Bioscience Reports*, vol. 37, n° 1, p. BSR20160184, https://doi.org/10.1042/BSR20160184.

Racine E., W. Boehlen, and M. Sample (2019), « Healthcare uses of artificial intelligence: Challenges and opportunities for growth », *Healthcare Management Forum*, vol. 32, pp. 272–275, https://doi.org/10.1177/0840470419843831.

Rappaport S. M. (2016), « Genetic Factors Are Not the Major Causes of Chronic Diseases », *PLoS ONE*, vol. 11, p. e0154387, https://doi.org/10.1371/journal.pone.0154387.

Robinson M. and G. Glusman (2018), « Genotype Fingerprints Enable Fast and Private Comparison of Genetic Testing Results for Research and Direct-to-Consumer Applications », *Genes*, vol. 9, p. 481, https://doi.org/10.3390/genes9100481.

Rocher L., J. M. Hendrickx, and Y.-A. de Montjoye (2019), « Estimating the success of re-identifications in incomplete datasets using generative models », *Nature Communications*, vol. 10, p. 3069, https://doi.org/10.1038/s41467-019-10933-3.

Ross C. and I. Swetlitz (2018), « IBM's Watson recommended "unsafe and incorrect" cancer treatments », *in STAT*, https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (Accessed 10 Nov 2019).

Rubinstein I. S. and W. Hartzog (2016), « Anonymization and Risk », *Washington Law Revue*, vol. 91, p. 59.

Ruby J. G., K. M. Wright, K. A. Rand *et al*. (2018), « Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating », *Genetics*, vol. 210, pp. 1109–1124, https://doi.org/10.1534/genetics.118.301613.

Rumelhart D. E., G. E. Hinton, and R. J. Williams (1986), « Learning representations by back-propagating errors », *Nature*, vol. 323, pp. 533–536, https://doi.org/10.1038/323533a0.

Schwarzinger M., B. G. Pollock, O. S. M. Hasan *et al*. (2018), « Contribution of alcohol use disorders to the burden of dementia in France 2008-13: a nationwide retrospective cohort study », The *Lancet Public Health*, vol. 3, pp. e124–e132, https://doi.org/10.1016/S2468-2667(18)30022-7.

Silver D., T. Hubert, J. Schrittwieser *et al*. (2017), « Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm », arXiv:171201815 [cs].

Topol E. J. (2019), « High-performance medicine: the convergence of human and artificial intelligence », *Nature Medicine*, vol. 25, pp. 44–56, https://doi.org/10.1038/s41591-018-0300-7.

Udler M. S., M. I. McCarthy, J. C. Florez, and A. Mahajan (2019), « Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine », *Endocrine Reviews*, vol. 40, pp. 1500–1520, https://doi.org/10.1210/er.2019-00088.

Université de Montréal (2018), « La déclaration de Montréal pour le développement responsable de l'intelligence artificielle ».

Vayena E., A. Blasimme, and I. G. Cohen (2018), « Machine learning in medicine: Addressing ethical challenges », *PLoS Med*, vol. 15, p. e1002689, https://doi.org/10.1371/journal.pmed.1002689

Villani C., M. Schoeunauer, Y. Bonnet *et al*. (2018), « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », La Documentation Française.

Vyas D. A., L. G. Eisenstein, and D. S. Jones (2020), « Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms », *The New England Journal of Medicine*, vol. 383, pp. 874–882, https://doi.org/10.1056/NEJMms2004740.

Wei Z., W. Wang, J. Bradfield *et al*. (2013), « Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease », *The American Journal of Human Genetics*, vol. 92, pp. 1008–1012, https://doi.org/10.1016/j.ajhg.2013.05.002.

Wiens J., S. Saria, M. Sendak *et al*. (2019), « Do no harm: a roadmap for responsible machine learning for health care », *Nature Medicine*, vol. 25, pp. 1337–1340, https://doi.org/10.1038/s41591-019-0548-6.

Williams S. A., M. Kivimaki, C. Langenberg *et al*. (2019), « Plasma protein patterns as comprehensive indicators of health », *Nature Medicine*, vol. 25, pp. 1851–1857, https://doi.org/10.1038/s41591-019-0665-2.

Wright K. M., K. A. Rand, A. Kermany *et al*. (2019), « A Prospective Analysis of Genetic Variants Associated with Human Lifespan », *G3*, vol. 9, pp. 2863–2878, https://doi.org/10.1534/g3.119.400448.

Yala A., C. Lehman, T. Schuster *et al*. (2019), « A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction », *Radiology*, vol. 292, pp. 60–66, https://doi.org/10.1148/radiol.2019182716.

Zins M., S. Bonenfant, M. Carton *et al*. (2010), « The CONSTANCES cohort: an open epidemiological laboratory », *BMC Public Health*, vol. 10, p. 479, https://doi.org/10.1186/1471-2458-10-479.

Žliobaitė I. (2017), « Measuring discrimination in algorithmic decision making », *Data Mining and Knowledge Discovery*, vol. 31, pp. 1060–1089, https://doi.org/10.1007/s10618-017-0506-1.

Annexes

Annexe 1 – Area under the curve ou AUC

Tout modèle ou algorithme d'apprentissage produit, pour chaque individu sur l'échantillon test indépendant, une prévision, sous la forme d'une probabilité $p(x_i)$ entre 0 et 1, d'avoir affaire à un mauvais payeur, de satisfaire à un emploi, de développer une pathologie... Cette probabilité est ensuite comparée à une valeur seuil *s* (par défaut 0,5) pour une prise de décision binaire. La comparaison entre les décisions, dépendant de *s*, et les vraies valeurs observées sur l'échantillon test conduit à la construction d'une table (tableau 2) de contingence ou matrice de confusion.

Tableau 2 – *Table de contingence croisant la prévision avec la valeur observée sur l'échantillon test indépendant ; toutes les quantités dépendent de la valeur seuil s choisie a priori.*

Prévision	**Observation**		
	Oui	**Non**	**Marge**
Oui	*a(s)*	*b(s)*	*a* + *b*
Non	*c(s)*	*d(s)*	*c* + *d*
Marge	*a* + *c*	*b* + *d*	*n* = *a* + *b* + *c* + *d*

Dans cette matrice, *a(s)* désigne le nombre de bonnes décisions ou vrais positifs, *d(s)* le nombre de vrais négatifs, *c(s)* le nombre de faux négatifs et *b(s)* le nombre de faux positifs. Le taux d'erreur est simplement défini par Terr = (*b* + *c*)/*n* mais est généralement insuffisant pour apprécier la qualité d'une prévision surtout si les classes sont déséquilibrées. De très nombreux indicateurs ont été définis dont la *sensibilité* ou taux de vrais positifs : TPR = *a*/(*a* + *c*) ; la *spécificité* ou taux de vrais négatifs : TNR = *d*/(*b* + *c*) ; le taux de faux positifs : FPR = *b*/(*b* + *c*) qui est encore *un moins la spécificité*. En faisant varier le seuil *s*, il est possible de tracer la courbe ROC (*receiver operating caracteristic*) exprimant TPR en fonction de FPR, dont la figure 1 donne un exemple. Plus la courbe se rapproche du cadre supérieur avec une croissance rapide, meilleure est la prévision avec la détection nette (valeur élevée de *s*) d'une grande proportion de vrais positifs en limitant la part de faux positifs ; l'AUC (entre 0,5 et 1) est l'aire délimitée par cette courbe. Si la courbe est proche de la diagonale (AUC = 0,5), la prévision n'est qu'un tirage à pile ou face. Cet indicateur permet de comparer les qualités de prévision de différents modèles ou algorithmes, celle-ci est jugée « bonne » au-delà de 0,8, excellente au-delà de 0,9.

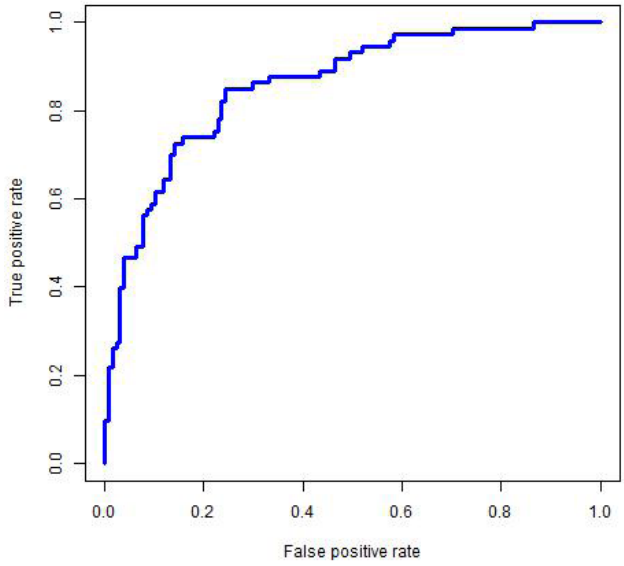


Figure 1 – *Exemple de courbe ROC pour la prévision de la variable binaire : dépassement du seuil de pic d'ozone. Les deux taux de vrais et faux positifs sont représentés en fonction de la valeur décroissante du seuil s de décision. Cette courbe aide à choisir, choix politique, la valeur de s conduisant à une détection d'un taux raisonnable de vrais positifs (pollution effective) pour la santé publique au regard d'un taux de faux positifs (absence de pollution) économiquement admissible.*

Annexe 2 – Analyse pangénomique des maladies multifactorielles

L'analyse ou médecine génomique appliquée aux maladies multigéniques, multifactorielles, généralement chroniques (cardiovasculaires, obésité, diabète, certains cancers…) et, plus idéalement encore, à l'étude de la durée de vie (Wright *et al*., 2019), soulève un ensemble de questions sur la pertinence des résultats, leur intérêt thérapeutique et leurs finalités au regard des risques encourus. Rappaport (2016) montre que les facteurs génétiques ne sont pas les facteurs majeurs des maladies chroniques. Pujol (2019) explique que, contrairement aux maladies monogéniques, la pénétrance d'une combinaison d'un nombre même important de variants génétiques est très faible, généralement quelques pourcents. Patron *et al*. (2019) proposent un outil pour estimer la qualité prédictive d'études pangénomiques publiées. Appliqué à 569 études, leur outil montre que très peu produisent une AUC (cf. définition annexe 1) plus grande que 0,75 ; la prédictibilité reste très nettement inférieure à celle obtenue avec les seules mesures cliniques classiques.

Ho *et al*. (2019) font la promotion de l'utilisation de l'apprentissage automatique, avec des maladresses sur sa présentation, par rapport à des facteurs de risques linéaires basés sur des variants jugés significativement influents ; ils citent quelques travaux plus prometteurs. Wei *et al*. (2013) obtiennent une qualité prédictive raisonnable (AUC de 0,83) de maladies inflammatoires de l'intestin en utilisant une régression (linéaire) logistique pénalisée (Lasso) pour opérer la sélection de quatre à cinq cents SNP après une pré-sélection (tests multiples) de 10800 parmi près de 179000. Lopez *et al*. (2018) utilisent l'algorithme des forêts aléatoires pour évaluer le risque d'occurrence du diabète de type 2 (T2D) pour aboutir à une AUC de 0,85. Les données initiales de ces deux analyses ont été abondamment nettoyées et sélectionnées pour produire des résultats tout à fait honorables sur des données largement prétraitées, mais, comme le rappelle Liu *et al*. (2019b), sont-ils reproductibles ? Qu'en serait-il sur de nouvelles données réelles, en général moins propres et pas nécessairement issues du même protocole

technologique ? Malgré ce qui est avancé, Mieth *et al*. (2016) s'intéressent à l'objectif de sélection de SNP influents mais pas à une prévision. Montañez *et al*. (2018) obtiennent pour la prévision de l'obésité une qualité (AUC de 0,99) qui éveille la suspicion. Cette qualité est obtenue en considérant 2465 SNP, issus d'un premier filtrage (tests) de 241000, alimentant un réseau de neurones avec deux couches cachées entraîné sur seulement 1200 individus. Ces quelques chiffres laissent présager une **situation de sur-apprentissage** mais qui ne peut être infirmée ou confirmée avec certitude sans disposer du code de l'analyse. Néanmoins, le descriptif succinct de la démarche confirme ce doute. Elle suit en détail les étapes du début du tutoriel du logiciel H2O[22] utilisé qui introduit une confusion dans les rôles des échantillons d'apprentissage, validation et test. La démarche semble donc manquer de rigueur et nous pouvons douter de la reproductibilité de la prévision sur un autre échantillon complètement indépendant.

Par ailleurs Lopez *et al*. (2018) regardent l'amélioration de la prévision du T2D par l'ajout de variables cliniques en ne respectant pas les standards cliniques comme les valeurs du glucose ou de l'insuline sanguin à jeun. Une revue plus récente et mieux documentée sur la même pathologie (Udler *et al*., 2019) conduit à des résultats plus nuancés dans l'utilisation de scores polygéniques : des valeurs d'AUC systématiquement plus faibles que celles fournies par des variables cliniques et une amélioration non significative de ces dernières lorsque les variables génomiques sont ajoutées. Lors du suivi d'une autre cohorte, Kraege *et al*. (2020) obtiennent un AUC de 0,85 très concurrentiel sur la base d'un seul score clinico-biologique (Kahn, 2009). Une des questions est donc de savoir si une analyse génétique, encore relativement coûteuse même si une seule suffit pour la vie, apporte une prévision du risque significativement plus précise qu'une analyse longitudinale des variables cliniques (Alaa *et al*., 2019). Il ne faut pas pour autant nier la part génétique de certaines maladies mais pour les principales maladies, notamment cardio-vasculaires (Do *et al*., 2012) ou T2D (Gim *et al*., 2017), la connaissance de l'historique familiale fait tout aussi bien. De plus, l'absence de prise en compte des conditions environnementales ou des styles de vie dans l'ensemble des études peut être une source de biais ou plutôt de confusion additionnelle en plus des biais ethniques déjà mentionnés. Une proximité phylogénétique (SNP ou historique familiale) peut être corrélée à une proximité géographique, sociologique donc de style de vie. Sans précision à ce sujet il est alors difficile de faire la part des choses entre les effets respectifs génétiques ou environnementaux et même leurs interactions potentielles.

Ceci n'empêche pas la société *Calico* (*California Life Company filiale d'Alphabet*) de continuer à financer l'étude de la détection de variants génétiques influençant la durée de vie par des modèles de Cox (Wright *et al*., 2019) appliqués aux bases pangénomiques, alors que Ruby *et al*. (2018) estiment à moins de 10 % la part génétique dans la durée de vie humaine.

L'étape suivante de cette démarche devrait déployer des algorithmes d'apprentissage statistique sur des cohortes très volumineuses associant données longitudinales environnementales, biologiques et cliniques, ainsi que génomiques. L'étude des possibles interactions entre variables environnementales, dont les effets sont connus, et génétiques nécessiterait, en raison de l'extrême complexité des phénomènes en jeu, des volumes de données considérables sans pour autant être sûr, à ce jour, de la pertinence des résultats attendus. Est-ce socialement acceptable compte tenu des risques encourus ?

Après cet aperçu illustratif et partiel, donc sans doute partial, d'une littérature très volumineuse, il est difficile d'établir une synthèse claire mettant en évidence un intérêt public substantiel qu'apporterait l'intégration des données génomiques dans la recherche sur les maladies multifactorielles chroniques.

22. http://docs.h2o.ai/h2o-tutorials/latest-stable/tutorials/deeplearning/index.html

Résumé succinct de la progression épistémologique récente de la recherche sur les maladies multifactorielles visant à intégrer ou associer différents types de données et méthodes d'analyse :

- des études épidémiologiques, basées sur des cohortes intégrant des données longitudinales, mettent en évidence, par des modèles statistiques, l'importance des facteurs environnementaux et des modes de vie ;
- des études pangénomiques, toujours basées sur des modèles statistiques, révèlent des listes importantes de variants génétiques, chacun de pénétrance faible, susceptibles d'influer sur le risque de maladie ;
- la prise en compte de ces facteurs génétiques n'améliore que très marginalement les qualités prédictives des scores cliniques, alors que des dosages de protéines (Williams *et al.*, 2019) conduisent eux à des résultats très encourageants ;
- l'utilisation d'algorithmes d'apprentissage statistique sur données génomiques, ou génomiques et cliniques, pour établir ces prévisions n'apporte pas de résultats plus probants ou soulève des questions quant à la rigueur de la mise en œuvre de ces outils sophistiqués, efficaces, mais excessivement sensibles à toute forme de biais : représentativité et nettoyage des données, gestion des échantillons d'entraînement, validation et test, validation sur un échantillon test réellement indépendant. Le principal risque est la construction de résultats spécifiques aux données étudiées, des artefacts non reproductibles. Ainsi, la notion de biais de sélection a déjà été identifiée par Ambroise et McLachlan (2002) dans les études transcriptomiques lorsqu'un algorithme de prévision est entraîné sur une sélection de variables sans intégrer la sélection opérée dans l'estimation de l'erreur de prévision. Mieth *et al.* (2016) induisent le même type de biais en présélectionnant une liste de SNP à l'aide d'un algorithme de SVM linéaire avant d'opérer la sélection classique par tests multiples mais avec une correction de Bonferroni inadaptée car calculée sur le seul nombre de tests.