Projet ANR

Deep Learning, multi-omiques et médecine personnalisée pour la prédiction de survie de patients à risque

Gabriel Cretin

7 janvier 2019

Table des matières

1 Contexte, positionnement et objectifs de la proposition

1.1 Objectifs et hypothèses scientifiques

Les cancers sont des maladies malheureusement de plus en plus communes dans le monde. Dans l'ensemble, ils font partie des principales causes de décès au niveau mondial voir même national pour certains pays, et leur incidence augmente avec le vieillissement de la population [1]. Ils sont également particulièrement craint en raison de leur létalité, de leurs symptômes et des thérapies (chimiothérapie, radiothérapie, chirurgie, etc) souvent assez toxiques et lourdes (nombreux effets secondaires indésirables parfois défigurants) utilisées pour les traiter. Depuis maintenant plusieurs années, un nouveau terme est apparu dans le domaine de la médecine, il s'agit de la médecine de précision. Ce terme est utilisé pour décrire un traitement individualisé qui englobe l'utilisation à la fois de nouveaux diagnostics et traitements, ciblés sur les besoins d'un patient en fonction de ses propres caractéristiques génétiques, phénotypiques, psychosociales et de ses biomarqueurs [2]. En particulier, des avancées dans des domaines tels que l'épigénétique, la protéomique, la métabolomique, etc., convergent avec l'informatique et d'autres technologies de pointe de manière à élargir rapidement la portée de ce domaine, qui est très prometteur pour faire avancer la recherche dans le domaine de la lutte contre le cancer par exemple.

La médecine personnalisée, individualisée, ou de précision (MP, ces termes sont généralement utilisés de façon interchangeable et désignent la même chose) est donc un pans de la médecine qui a le potentiel d'adapter un traitement avec la meilleure réponse et la marge de sécurité la plus élevée possible pour assurer de meilleurs soins aux patients [3]. En permettant à chaque patient de recevoir des diagnostics plus tôt, des évaluations des risques et des traitements optimaux, la MP est prometteuse pour améliorer les soins de santé tout en réduisant les coûts. De plus, avec la généralisation des dossier médicaux électroniques ainsi que de l'augmentation de la précision des mesures de constantes biologiques avec notamment les techniques de nouvelle génération à haut débit pour le séquençage par exemple, il y a une croissance formidable de la quantité de données potentiellement exploitables automatiquement et intelligemment grâce à l'informatique et à des outils d'apprentissage automatique aujourd'hui assez répandus dans des domaines tels que l'intelligence artificielle et les nouvelles technologies de l'information avec le machine learning et le deep learning.

Ces dossiers médicaux électroniques permettent de compiler quantités d'informations diverses et variées sur le profil phénotypique des patients, fournissant alors un accès privilégié à des données personnalisées, précises et par là même très complètes et utiles, il s'agit d'une mine d'or encore trop inexploitée dans le domaine de la médecine. Les domaines d'étude "omiques" (génomique, transcriptomique, protéomique, mé-

tabolomique, etc.) peuvent être corrélées à ces profiles phénotypiques pour mieux comprendre et appréhender les réponses aux traitements et la toxicité. Ceci en rajoutant également un aspect temporel dans l'acquisition de ces mesures afin de dresser un profil dans le temps de l'évolution de certaines constantes biologiques plus particulièrement.

La combinaison de données issues de dossiers médicaux personnalisés avec des données multi-omiques mesurées sur une période donnée à l'aide de méthodes d'apprentissage automatique comme du Deep Learning, pourrait donc générer des données de haute-qualité pour la médecine de précision. Cela offrirait également la possibilité de prédire certains phénotypes ou réponses à des traitements contre le cancer par exemple. En effet, ce raisonnement s'appuie sur le constat que chaque patient présente un ensemble spécifique d'anomalies moléculaires responsables de sa maladie ou en corrélation avec la réponse au traitement et les résultats cliniques [4]. Avec une quantité de données assez importante, précises et diversifiée sur les profiles, il devrait être possible de cibler les spécificités de ces derniers grâce à un algorithme d'apprentissage profond.

La médecine personnalisée n'est pas simplement un enjeux de santé public, elle a également un impact sociétal important et un impact certain et non négligeable sur l'industrie pharmaceutique. En France, l'incidence des hospitalisations dues à des accidents médicamenteux est estimée à 3.2% par an, pour un coût global annuel d'environ 320 millions d'euros [5]. La MP offre alors l'opportunité de développer des traitements qui ciblent des groupes de patients qui ne répondent pas de manière attendue aux traitements usuels, et pour lesquels le système de santé traditionnel n'a pas fonctionné non plus. L'apport de l'apprentissage automatique profond sur des données médicales et phénotypiques personnalisées et mesurées sur une période de temps donnée est alors déterminant pour non seulement baisser les coûts de développement pour l'industrie pharmaceutique, mais dans le même temps augmenter les chances de développer des traitements vraiment utiles car ciblés pour chaque patient.

Tout l'enjeu de ce challenge ambitieux est dans un premier temps l'obtention des données. Il existe des bases de données de cohortes pour lesquelles des mesures multi-omiques temporelles ont déjà été effectuées. D'autres données devront être générées pour compléter le jeux de données, car la clé de voûte de l'apprentissage automatique profond est le volume de données disponible. Plus il y a de données issues de patients différents, meilleur est l'apprentissage car meilleur est le signal statistique derrière. La diversité des données permettra alors à l'apprentissage automatique de mieux déterminer les spécificités inhérentes à chaque profile biologique, c'est-à-dire à chaque patient.

1.2 État de l'art

1.2.1 La médecine personnalisée

La médecine personnalisée, la médecine de précision ou aussi appelée théranostique (contraction de thérapie et diagnostic) est un modèle médical qui divise les personnes en différents groupes. Les décisions, pratiques, interventions et/ou produits médicaux sont adaptés au patient en fonction de la réponse ou du risque de maladie prédit [6]. Elle est construite sur une caractérisation fine du patient notamment basée sur une l'analyse moléculaire et les signatures génétiques afin d'être plus prédictive dans le but d'optimiser les chances de guérison, d'efficacité des médicaments (pharmacogénétique) et d'améliorer la gestion de la maladie en identifiant mieux les prédispositions à une maladie d'un patient, aux éventuels effets secondaires d'un traitement, etc.

Des facteurs génétiques affectant la pharmacocinétique et la pharmacodynamie des médicaments expliquent en partie la variabilité interindividuelle [5].

La MP a depuis près d'une dizaine d'années pris un essor certain notamment avec les avancées majeurs qui ont été faites en parallèle en informatique (puissance de calcul), bioinformatique (algorithmes, outils, plates-formes dédiées) et dans les domaines des omiques, en particulier en génomique, avec le séquençage nouvelle génération et toutes les technologies de haut débit qui ont pu se développer en transcriptomique et métabolomique (plus haute résolution des spectromètres de masse).

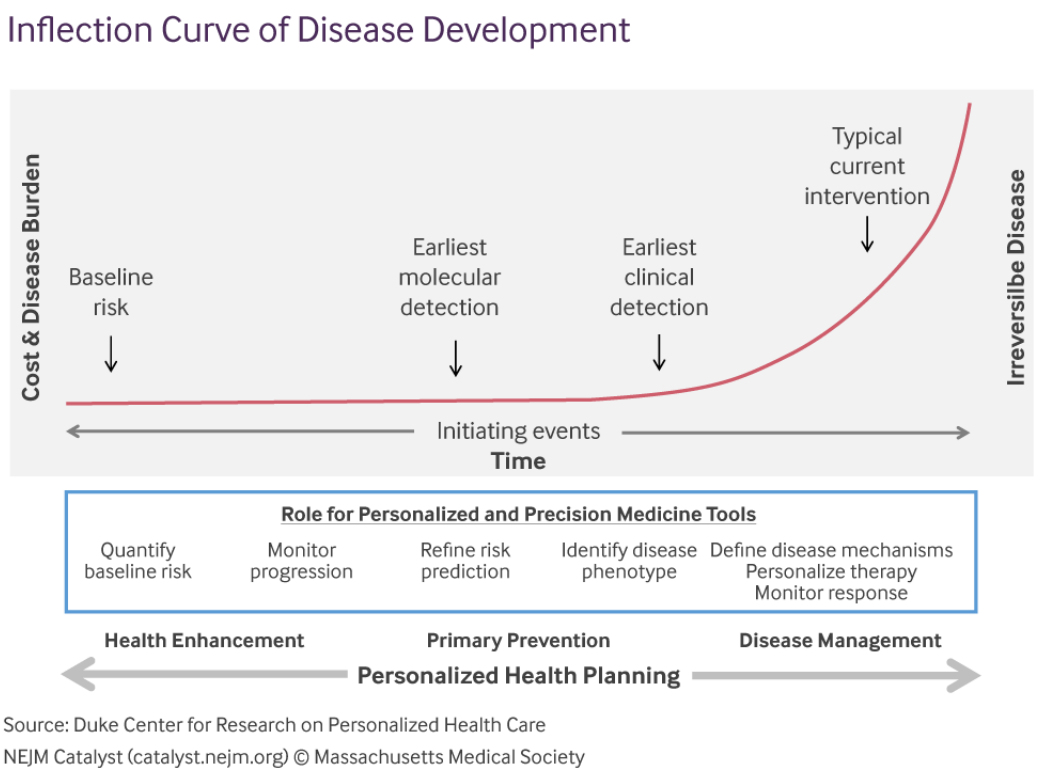

FIGURE 1 –

La figure 1 montre les étapes de l'évolution d'une maladie en mettant l'accent sur l'impact qu'a de manière générale l'attente, ou plutôt l'inaction face à elle. Cette figure parle d'elle-même, et montre de façon claire l'impact qu'aurait la proaction vs réaction et l'importance de devoir prendre en charge assez tôt la maladie. Cela implique bien évidement d'avoir les moyens à disposition pour détecter les signes, biomarqueurs, anomalies génétiques, etc.

Il a été montré que le risque de maladie peut être estimé à l'aide de la séquence d'un génome entier et par un suivi régulier des états de santé avec l'aide d'un profile omique personnel intégré (iPOP) [7]. La richesse des informations fournies par un profil iPOP longitudinal détaillé a révélé une complexité moléculaire inattendue, qui a présenté des changements dynamiques au cours d'états sains et malades, et a permis de mieux comprendre de nombreux processus biologiques. Le profilage détaillé de l'omique associé au séquençage du génome peut fournir des informations moléculaires et physiologiques d'importance médicale.

L'étude iPOP [7] a :

— Établie une cohorte d'environ 100 personnes consentantes à risque de diabète. Les échantillons ont été prélevés fréquemment (à intervalles de 1 à 4 jours) lors d'infections et d'autres états de stress, et moins fréquemment (tous les 3 mois) au cours de périodes saines, avec un minimum de 27 points

d'échantillonnage dans le temps, par sujet.

— Effectuée une analyse omique des échantillons de patients longitudinaux et des échantillons de leurs microbiomes. Le contenu génétique, le transcriptome et le protéome du microbiome ont été déterminés dans des échantillons fécaux et nasaux (pour les populations de microbiomes viraux et endogènes). Les données génomiques, transcriptomiques et protéomiques des patients ont été analysées simultanément à partir du sang (PBMC (cellules mononucléées du sang périphérique) pour le génome et le transcriptome ; plasma pour le protéome). Les métabolites du patient et du microbiome combinés ont été quant à eux analysés dans le sérum et l'urine. Les anticorps anti-virome ont également été analysés.

— Analysée et intégrée les données du patient et celles du microbiome. Les différentes données de microbiome viral et fécal (génome, transcriptome et protéome) ont été analysées individuellement et de manière combinée afin de déterminer les voies dynamiques qui changent au cours de l'infection virale. En outre, une corrélation entre les profils temporels des données du microbiome et ceux des patients a été établie afin d'obtenir une vision sans précédent des changements globaux microbiome-hôte survenant au cours d'une infection virale.

Dans l'ensemble, l'étude longitudinale a révélée des changements globaux dans le microbiome des patients à un niveau de détail sans précédent et nous permettre d'identifier les molécules et les voies qui changent au cours des infections virales ainsi que de l'apparition et de la progression du diabète. Néanmoins, nous tenterons dans ce projet-ci de voir si ce genre de découvertes ne peut pas également avoir un impact sur la découverte de biomarkeurs pour le cancer par example.

1.2.2 Machine / Deep learning

L'apprentissage automatique (*Machine Learning*, ML) est omniprésent dans les produits de tous les jours tels que les recherches sur Internet, les filtres anti-spam, les recommandations de produits, la classification des images et la reconnaissance vocale. De nouvelles approches d'automatisation hautement intégrées telles que l'Internet des objets convergent également avec les méthodologies ML. De nombreuses approches incorporent des architectures de réseaux de neurones artificiels complexes et sont collectivement appelées applications d'apprentissage en profondeur (DL). Ces méthodes ont démontré leur capacité à représenter et à apprendre des relations prévisibles dans de nombreuses formes de données et sont prometteuses pour transformer le futur de la recherche et des applications en omique en médecine de précision. Les données omiques et les dossiers de santé électroniques posent des défis considérables au DL. De nombreux facteurs entrent en jeu et sont en cause, tels que le faible rapport signal sur bruit, la variance analytique et les exigences complexes d'intégration des données. De plus, le domaine des omiques comprend une multitude de sous domaines de

recherche tels que la génomique, la transcriptomique, la protéomique, l'interactomique, la métabolomique, la phénomique et la pharmacogénomique, pour ne citer que quelques-uns. Chacun de ces domaines peut également comporter de nombreux sous-domaines, nécessitant chacun une spécialisation plus poussée dans les approches analytiques et (bio)informatiques. De surcroît, l'ampleur de la génération de données omiques met à l'épreuve la capacité des chercheurs à intégrer et à modéliser des données souvent saturées en bruit, complexes et de grande dimension. Le *deep learning* est un sous-domaine du *machine learning* qui s'est révélé être une approche puissante qui peut à la fois coder et modéliser de nombreuses formes de données complexes comme par exemple de type numériques, texte, audio et image, à la fois en mode supervisé (identification de biomarqueurs) et non supervisé (détection d'anomalies). En effet, il a déjà été démontré que les modèles DL pouvaient à la fois améliorer la facilité de codage des données et les performances des modèles prédictifs par rapport aux approches alternatives [8].

L'apprentissage automatique et/ou profond a déjà été utilisé dans le domaine de la radio-oncologie par exemple afin d'essayer de prédire les résultats potentiel de traitements [4]. L'étude s'est appuyée sur l'opportunité qu'offre la généralisation des dossiers de santé électroniques afin de générer des profiles phénotypiques complets. Elle a également montré et décrit les caractéristiques essentielles qui doivent être prises en compte et intégrées dans un modèle de prédiction issu de machine/deep learning. Les mêmes challenges peuvent donc s'appliquer pour toute autre étude qui s'intéresse à l'intégration de données multi-omiques.

Plusieurs méthodologies et outils de bioinformatique existent pour l'intégration de différents types de données issues des multi-omiques. Elles utilisent l'exploration de données (*data mining*) et des algorithmes de prédiction. Pour cela, les approches *Machine Learning and Systems Genomics* (MLSG) permettent une interprétation bien plus cohérente des relations phénotype-génotype qu'une analyse se basant uniquement sur un seul type de données. Il est alors nécessaire d'utiliser des outils spécialisés afin d'être en mesure de générer des prédictions à partir d'un phénotype donné en utilisant des données multi-omiques nouvelle génération.

Les méthodes développées dans le cadre d'applications MLSG (*Machine Learning and Systems Genomic*) englobent : l'intégration basée sur les modèles (MBI, *Model-based Integration*), l'intégration basée sur la concaténation (CBI, *Concatenation-Based Integration*) et enfin l'intégration basée sur des transformations (TBI, *Transformation-Based Integration* [9].

Dans un premier temps il est nécessaire d'avoir accès à des données personnalisées d'une cohorte de patients sains [7] ainsi qu'à une base de données facilement accessible et utilisable par des outils d'apprentissage automatique de données issues d'études du génome du cancer [10].

1.3 Originalité et pertinence par rapport à l'état de l'art

L'originalité de ce projet se trouve dans son caractère ambitieux et innovateur par sa volonté de combiner des travaux de grande ampleur, sur des cohortes et avec de multiples données, avec des nouvelles technologies habituellement retrouvées dans des domaines tout à fait différents de l'information et de la communication, afin de proposer une réponse à un enjeu sociétal, économique et tout simplement de santé publique. En effet, ce projet se propose non pas seulement d'intégrer des donnée multi-omiques, ce qui a déjà été réalisé, mais le rajout d'une dimension temporelle et personnalisée aux données, suivi de l'application d'un apprentissage automatique profond sur ces données par des algorithmes de deep learning performant pour la prédiction de survie de patients à risque pour le cancer et de prédiction de réaction à des traitements le cas échéant.

2 Organisation du projet et moyens mis en œuvre

2.1 Coordinateur scientifique

Le projet sera conduit par un directeur de recherche. Tous les postes et coûts associés sont représentés dans le tableau 3

2.2 Consortium

Plusieurs axes de compétences sont nécessaires dans le cadre de ce projet. Il est nécessaire dans un premier temps d'avoir une expertise biologique avec une plate-forme dédiée aux prélèvements biologiques. Puis une plate-forme bioinformatique qui se charge de traiter les échantillons pour réaliser les mesures des constantes omiques (GWAS, transcriptomique et métabolomique). Une autre et dernière équipe plutôt chargé développement en bioinformatique est quant à elle axée sur le développement logiciel de méthodes et outils pour effectuer le machine learning ou le deep learning sur la grande masse de données.

2.2.1 Plate-forme médicale

La plate-forme médicale sera chargée de réaliser les différents prélèvements biologiques sur les patients. Suite à une campagne de recrutement de patients auprès d'organismes spécialisés, l'équipe devra suivre un protocole précis pour chacun des prélèvements à effectuer.

2.2.2 Plate-forme "omiques"

Cette plate-forme (potentiellement plusieurs partenaires répartis sur le territoire selon les spécialités), est chargée de réaliser les traitement des échantillons. Une étude GWAS sera effectuée, suivie d'une analyse

transcriptomique, puis métabolomique, afin de récolter les mêmes types de données que l'étude iPOP.

2.2.3 Plate-forme de bioinformatique

Une équipe de 4 personnes sera chargée de réaliser toutes les études statistiques, informatiques et bioinformatiques. A savoir, le pré-traitement statistique des données une fois générées ; sondage de la littérature pour chercher les méthodes de deep learning les plus adaptées à l'intégration de ce type de données, éventuellement en développer de nouvelles ; développement et mise en place du pipeline de traitement de toutes les données en portant une attention particulière à l'anonymisation et sécurité des données ; analyses bioinformatiques par la suite.

2.3 Calendrier prévisionnel

Echéancier: calendrier prévisionnel des tâches

Tâches	Année 1		Année 2		Année 3	
	S1	S2	S3	S4	S5	S6
Mesures omiques sur la nouvelle cohorte	18 mois					
Préparation + analyse des données omiques de iPOP	18 mois					
Développement des méthodes machine/deep learning se basant sur données iPOP	18 mois					
Fusion des données iPOP + nouvelle cohorte				4 mois		
Application du deep learning sur toutes les données					1 an	
Analyses bioinformatiques des résultats + optimisation des méthodes					1 an	

FIGURE 2 – Calendrier prévisionnel des tâches à effectuer pendant la durée totale du projet. Certaines tâches sont incompressibles, comme la mise en place de la cohorte

La figure 2 représente le schéma organisationnel prévisionnel du projet. En effet le projet se déroule en plusieurs étapes qui seront décrites dans cette partie.

2.3.1 Nouvelle cohorte

La première étape et non des moindres, est la mise en place et le prélèvement de mesures omiques sur une cohorte d'une centaine de personnes si possible. L'objectif est de faire des mesures omiques similaires à celles qui ont pu être effectuées dans le cadre de l'étude iPOP afin de pouvoir à terme les fusionner pour obtenir une quantité de données suffisante pour exercer du deep learning dessus.

Néanmoins, iPOP regroupe beaucoup de domaines omiques avec une grande quantité de données que nous ne pourrons pas reproduire sur notre cohorte pour des raisons évidentes de temps et de coût. Nous nous restreindrons à des mesures de GWAS, transcriptomiques, et métabolomiques. Des domaines des omiques où de nombreuses avancées technologiques depuis maintenant quelques années ont permis de développer de nouvelles techniques qui réduisent grandement les coups [11].
L'étude iPOP s'est basée sur des patients sains, mais a fait des prélèvements également lors de périodes de maladie pour les patients. Notre cohorte sera composée d'un tiers de patients atteints d'un cancer, peu importe le type de cancer et le stade, nous nous intéressons aux cellules cancéreuses de manière générale. Un autre tiers de patients dis à risque, c'est-à-dire des patients qui ont une hygiène de vie connue pour être potentiellement initiatrice de cancer (fumeurs, alcooliques, atteints d'obésité) ainsi que de patients qui ont des antécédents connus de cancer dans la famille par exemple. Le dernier tiers sera des patients sains. Ce panel de patient permet d'élargir au maximum les profiles phénotypiques, toujours dans l'objectif de diversifier les données pour maximiser les chances que le deep learning puisse apprendre correctement sur les données, et plus particulièrement si possible des biomarqueurs de cancers. Etant donné que les données doivent être similaires en terme de forme et de fond avec celles d'iPOP pour une fusion simplifiée, le même protocole sera appliqué. A savoir le suivi de patients pendant 14 mois, avec des prélèvements réguliers. Une période plus longue a été donnée à cette tâche pour prendre en compte le temps de recrutement des patients.

2.3.2 Préparation et analyses des données iPOP

Pendant que la cohorte est mise en place et que les prélèvements sont effectués, une autre équipe se chargera de commencer à regrouper et à préparer les données iPOP, qui elles, sont déjà prêtes : iPOP data
Cela permettra de se familiariser avec les données et de préparer l'arrivée des nouvelles données générées par la nouvelle cohorte. Un premier tri pourra être fait durant cette étape, c'est-à-dire que c'est durant cette étape que les équipes rendront compte de difficultés ou de modifictions à effectuer dans le planning.

2.3.3 Développement du Deep Learning

Toujours en attendant que les données de la cohorte apparaissent, l'équipe de bioinformaticiens peut en parallèle commencer à développer le programme de deep learning et les outils qui permettront d'analyser les futures données, en utilisant les données d'iPOP.

2.3.4 Fusion des données iPOP et cohorte

Cette étape est cruciale. C'est l'étape où toutes les données ont été prélevées pour la cohorte, et où l'équipe de bioinformaticiens les intègre aux autres donnes iPOP. Cette étape peut durer un certain temps

si le format des données est différent, s'il est nécessaire de transformer les données statistiquement parlant ou non.

2.3.5 Application du Deep Learning

Cette étape consiste à appliquer les méthodes de machine/deep learning sur l'ensemble des données fusionnées. Cette étape nécessitera une assez grande puissance de calcul pour pouvoir traiter la masse de données. Un cluster de calcul devra être utilisé.

2.4 Moyens mis en œuvre pour atteindre les objectifs

Chacune des étapes explicitées préalablement requièrent une expertise et des qualifications particulières en personnel. L'expertise en divisée en trois catégories : biomédicale, expérimentateur, bioinformatique. Les études GWAS, transcriptomiques et métabolomiques coûtent très cher étant donné les besoins en matériel et le nombre de patients à traiter durant les 14 mois.
Les coûts associés aux besoins en personnel sont décrits dans le tableau 3. Le coût total de ce projet ambitieux et novateur est estimé à 1020600€.

Budget associé au personnel				
Poste	Nombre Homme.mois	Coût Homme.mois (salaire chargé)	Nombre de personnes impliquées	Coût Total
Professeur	10	7500	1	75000
Ingénieur d'étude	24	3500	1	84000
Maître de conférences	18	4500	1	81000
Doctorants	36	2300	4	331200
Stagiaires	12	575	6	41400
Total				612600

Budget associé aux équipements et matériel		
	Nombre	Coût €
OMIQUES (GWAS + transcriptomique + métabolomique)	100	400000
Cluster	1	8000
Total		408000
TOTAL		1020600

FIGURE 3 – Budget à estimation large du projet. Ce budget est général et ne peut pas prendre en compte toutes les dépenses annexes ni le salaire exact des personnes impliquées.

3 Discussion & Conclusion

Pour aller plus loin

Pour aller plus loin dans la démarche et rajouter de l'information, il serait intéressant de pouvoir effectuer le même genre d'analyses en rajoutant des données externes, non pas omiques. Par exemple, les pays scandinaves et le Canada (avec l'Université de la Colombie-Britannique) ont récemment construit des *Population Data Centers* reliés aux systèmes d'information médico-sociaux existants, mis à disposition (à certaines conditions) des acteurs de la santé publique et de la recherche [12]. Ils offrent aux chercheurs l'accès à l'une des plus grandes collections au monde de données de santé, les services de santé et données sur la santé de la population. Ce système est issu de la *Population Data BC* (PopData), qui est une ressource multi-universitaire de données et éducative facilitant la recherche interdisciplinaire sur les déterminants de la santé humaine, du bien-être et du développement.

Pour la problématique de cette étude, à savoir le cancer, il serait intéressant de tirer des informations de la base de données *BC Cancer Registry Data* qui en est issue. En effet, cette dernière rapporte les informations sur tous les cancers diagnostiqués pour les résidents de Colombie-Britannique (BC). En Colombie-

Britannique, le cancer est une maladie à déclaration obligatoire. Les sources de données sont les rapports d'hématologie et de pathologie, les certificats de décès, les rapports d'hôpitaux et les centres de traitement du cancer. Les données récoltées sont d'abord assez générales (Identité, âge, date de naissance) puis plus précises sur le type de cancer, de tumeur, avec le code histologique de la tumeur, suivi de descriptions sur les tumeurs. Cela représente une mine d'or à exploiter pour notre étude. D'autant plus que les autorités estiment que le *BC Cancer Registry* couvre au moins 95% de tous les cas de cancer. Cette base de données inclue tous les cas dont le code postal est de Colombie-Britannique ou de province depuis 1985 à aujourd'hui. Elle exclu tous les cas bénins, ainsi que tous les cas en dehors du Canada. Ce type de données peut donner lieu à une investigation par des méthodes de fouille de données "data mining" qui serait une source d'information en plus à relier aux données de machine learning.

Cette plate-forme propose l'accès de la recherche à des données longitudinales anonymes et individualisées sur les 4,7 millions d'habitants de la Colombie-Britannique. De surcroît, ces données peuvent être liées les unes aux autres et à des ensembles de données externes, sur approbation du fournisseur de données. La mise en relation de données entre différents secteurs, tels que la santé, l'éducation, le développement de la petite enfance, le lieu de travail et l'environnement, permet de mieux comprendre l'interaction complexe des influences sur la santé, le bien-être et le développement humains. De telles recherches éclairent les décisions d'investissement et de décision en matière de santé.

Références

[1] Francis S. Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9) :793–795, 2 2015.

[2] J. Larry Jameson and Dan L. Longo. Precision Medicine—Personalized, Problematic, and Promising. *Obstetrical & Gynecological Survey*, 70(10) :612–614, 10 2015.

[3] F Randy Vogenberg, Carol Isaacson Barash, and Michael Pursel. Personalized medicine : part 1 : evolution and development into theranostics. *P & T : a peer-reviewed journal for formulary management*, 35(10) :560–76, 10 2010.

[4] Jean-Emmanuel Bibault, Philippe Giraud, and Anita Burgun. Big Data and machine learning in radiation oncology : State of the art and future prospects. *Cancer Letters*, 382(1) :110–117, 11 2016.

[5] D. Societe francaise de biologie clinique. and M.-A. Loriot. *Annales de biologie clinique.*, volume 62. Elsevier, 9 2004.

[6] Stratified, personalised or P4 medicine : a new direction for placing the patient at the centre of healthcare and health education (May 2015). Technical report, 2015.

[7] Rui Chen, George I. Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y.K. Lam, Rong Chen, Elana Miriami, Konrad J. Karczewski, Manoj Hariharan, Frederick E. Dewey, Yong Cheng, Michael J. Clark, Hogune Im, Lukas Habegger, Suganthi Balasubramanian, Maeve O'Huallachain, Joel T. Dudley, Sara Hillenmeyer, Rajini Haraksingh, Donald Sharon, Ghia Euskirchen, Phil Lacroute, Keith Bettinger, Alan P. Boyle, Maya Kasowski, Fabian Grubert, Scott Seki, Marco Garcia, Michelle Whirl-Carrillo, Mercedes Gallardo, Maria A. Blasco, Peter L. Greenberg, Phyllis Snyder, Teri E. Klein, Russ B. Altman, Atul J. Butte, Euan A. Ashley, Mark Gerstein, Kari C. Nadeau, Hua Tang, and Michael Snyder. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. *Cell*, 148(6) :1293–1307, 3 2012.

[8] Dmitry Grapov, Johannes Fahrmann, Kwanjeera Wanichthanarak, and Sakda Khoomrung. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. *OMICS : A Journal of Integrative Biology*, 22(10) :630–636, 10 2018.

[9] Eugene Lin and Hsien-Yuan Lane. Machine learning and systems genomics approaches for multi-omics data. *Biomarker Research*, 5(1) :2, 12 2017.

[10] Rasmus Krempel, Pranav Kulkarni, Annie Yim, Ulrich Lang, Bianca Habermann, and Peter Frommolt. Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). *BMC Bioinformatics*, 19(1) :156, 12 2018.

[11] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome : the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2) :286–302, 3 2018.

[12] Population Data BC | www.popdata.bc.ca.