Les *Big Data* et la prédiction en oncologie

Big Data and prediction in cancer

Julien Péron

Hospices civils de Lyon
Service d'oncologie médicale
165, chemin du Grand-Revoyet
69310 Pierre-Bénite
France
<julien.peron@chu-lyon.fr>

Remerciements et autres mentions :

Financement : aucun

Liens d'intérêts : l'auteur déclare ne pas avoir de lien d'intérêt.

Les *Big Data* ou méga-données n'ont pas de définition universelle, mais correspondent aux données dont le traitement dépasse les capacités des technologies courantes du fait de leur volume, vitesse (arrivée en continu de données), variété (plusieurs sources de données), véracité (erreurs plus ou moins structurées attendues) et valeur (mélange de données utiles et de données moins utiles) : ce sont les 5 « V » du *Big Data*.

Je vais m'intéresser dans cet éditorial aux méga-données en santé correspondant aux données générées par la numérisation croissante des données de santé (dossier patient informatisé, base du Système national d'information interrégimes de l'assurance maladie [SNIIRAM], base de données de recherche, résultats d'analyse de laboratoire d'imagerie ou de biologie [notamment génomique], objets connectés, et autres).

L'exploitation des méga-données a été rendue possible ces dernières années par la multiplication des sources d'information (objets connectés, numérisation, analyses génomiques), l'amélioration de la gestion des serveurs et du stockage à grande échelle (initiée par Google dans le but d'indexer le Web entier), et surtout l'amélioration des algorithmiques d'analyse (plus ou moins supervisés) permettant l'interrogation plus rapide des bases de données.

La valorisation des méga-données par les *data scientists* (experts en méga-données) n'est plus à démontrer, et ces données sont maintenant utilisées de façon courante dans de nombreux domaines comme la planification industrielle, le marketing, la surveillance, voire le domaine électoral. Pour prédire un phénomène, les outils de *machine learning* (dont l'intelligence artificielle) se montrent presque toujours plus performants que les méthodes de régression classiquement utilisées en modélisation statistique.

La santé est l'un des secteurs où les méga-données génèrent le plus d'attentes. L'amélioration de nos capacités prédictives (d'un risque de maladie, d'un pronostic, ou de l'efficacité d'un traitement) devrait permettre de perfectionner les diagnostics, de mieux personnaliser les traitements, d'organiser des dépistages individuels, et d'adapter constamment le système de soins aux besoins des patients. Ces progrès figurent, en tout cas, parmi les succès futurs annoncés.

À l'heure où de très nombreux domaines économiques ont vu leurs pratiques révolutionnées par des start-up basées sur une idée très innovante, ayant la capacité de se diffuser très rapidement, on tarde toujours à voir apparaître des exemples comparables en médecine. En effet, en observant les stratégies utilisées en 2018 pour développer un nouveau médicament, ou pour identifier et développer un nouveau biomarqueur, mon impression est que nos stratégies n'ont été que faiblement bousculées ces dernières décennies. Les formidables avancées de la chimie et de l'ingénierie nous permettent d'avoir accès à des médicaments plus ciblés, à des méthodes d'imagerie plus précises, et à des outils de diagnostic biologique d'une performance de plus en plus impressionnante. Néanmoins, la validation clinique (nécessaire !) des médicaments ou des biomarqueurs repose sur des paradigmes fondamentalement inchangés.

Un changement notable est que les essais cliniques modernes capturent de plus en plus souvent des informations de plusieurs types (génomiques, qualité de vie, imagerie, efficacité des traitements). Même si cela constitue

Tirés à part : J. Péron

Pour citer cet article : Péron J. Les *Big Data* et la prédiction en oncologie. *Innov Ther Oncol* 2018 ; 4 : 209-210. doi : 10.1684/ito.2018.0135

doi: 10.1684/ito.2018.0135

un progrès certain, le nombre de patients inclus dans les essais cliniques reste souvent limité, et les analyses de ces essais restent, par conséquent, le plus souvent conduites de façon standard. Pourtant, il semblerait logique, qu'après la démonstration de l'efficacité d'un médicament et sa mise sur le marché, les modèles permettant de prédire son efficacité en fonction des caractéristiques des patients soient constamment optimisés en utilisant des méga-bases de données cumulant toutes les informations disponibles sur les patients recevant ce traitement (incluant des données d'imagerie, de génomique, des informations cliniques, socioprofessionnelles et autres). Malheureusement, les exemples de telles analyses sont si rares qu'aucun exemple concret ayant réellement impacté les pratiques ne me vient à l'esprit.

Les outils informatiques permettant d'analyser ces données sont complexes, mais à la lumière des incroyables performances réalisées dans d'autres secteurs professionnels, l'explication ne vient probablement pas de problèmes techniques. En tout cas, cela ne devrait pas être le cas si les données adéquates étaient accessibles par des personnes expertes dans la gestion des méga-données.

Une autre explication possible est que le milieu médical doit impérativement conserver un compromis sans faille entre les innovations et leurs usages, afin de protéger nos patients, ce qui limite l'implémentation hasardeuse des nouvelles technologies. Cette exigence légitime des patients et de la société est très certainement une explication partielle. Néanmoins, il me semble peu probable que cette explication suffise, puisque nous parlons avant tout de modèles prédictifs qui permettraient d'améliorer l'utilisation de traitements déjà existants.

À mon sens, les deux principaux freins à l'exploitation des méga-données en santé à but prédictif sont :

- la façon dont sont actuellement utilisées et partagées les données de santé ;
- la façon dont sont structurées, dans les méga-bases, les données les plus pertinentes qui sont les données cliniques, biologiques et d'imagerie.

En effet dans un monde idéal, et pour imaginer que la connaissance avance le plus efficacement possible, il faudrait que des laboratoires de recherche qualifiés, composés de *data scientists*, de cogniticiens, de statisticiens et de médecins, puissent avoir accès à des données pseudo-anonymisées mais individuelles de patients, leur permettant de connaître les caractéristiques cliniques et biologiques de leurs maladies, leurs pronostics, et éventuellement leurs données comportementales et génétiques. De plus, l'accès aux données devrait être le plus ouvert possible afin de permettre d'évaluer la reproductibilité des analyses.

Ce partage d'information pose évidemment la question du consentement et de l'information des patients, ainsi que la question de leur identifiabilité à partir des données, même pseudo-anonymisées. Néanmoins de nombreux représentants de patients plaident en faveur de plus de partage de données entre chercheurs, et de plus de collaborations. Donc même si le cadre légal du partage de données nécessite d'être bien réfléchi, il ne semble pas que les patients ou les médecins y soient opposés.

En réalité, de nombreuses initiatives sont en cours pour structurer et constituer des méga-bases de données de santé. La France fait d'ailleurs plutôt mieux que la moyenne puisqu'il existe dans notre pays plusieurs centaines de bases de données publiques dans le domaine de la santé. La plus riche des bases médico-administratives est celle du SNIIRAM, dont les caractéristiques de notre modèle social assurent la richesse et la quasi-exhaustivité.

Néanmoins afin d'être rendues plus pertinentes, il est nécessaire de croiser ces bases avec des bases cliniques ou biologiques. Là encore, les initiatives françaises comme le programme ESMÉ (épidémio-stratégie médico-économique) des centres anticancéreux sont à louer. Cependant, la possibilité d'utiliser de façon fiable ces bases clinico-biologiques repose sur la structuration des données cliniques et biologiques : les données accessibles doivent être, autant que possible, structurées de telle façon que les données manquantes soient rares, et surtout que les données manquantes le soient de façon aléatoire. Si ces deux paramètres ne sont pas présents, la meilleure analyse de *machine learning* sera à haut risque de conclusion erronée. Le retour d'expérience de ceux qui développent et utilisent l'intelligence artificielle est que cette intelligence est plus collective qu'artificielle. Elle repose donc plus sur la capacité à mobiliser des humains pour faire apprendre à la machine que l'inverse. Il n'existe malheureusement pas, pour l'instant, de façon homogène de structurer les informations contenues dans nos dossiers patients. Les fiches de réunions de concertations pluridisciplinaires, dont la structuration minimale est une requête de l'Institut national du cancer (INCa), sont une magnifique opportunité de réaliser une structuration la plus homogène et la plus informative possible des données cliniques de nos patients.

L'objectif de cet éditorial n'est certainement pas de critiquer les institutions et les individus qui progressivement pavent le chemin du partage et de la valorisation des méga-données de santé. Mon but est de pointer le fait que chaque médecin devrait, dans la mesure de ses possibilités, œuvrer à faciliter le partage (légal) des données de santé et à structurer les informations sur ces patients. C'est à mon sens l'effort collectif nécessaire afin de pouvoir, nous aussi, tirer le plein bénéfice des progrès des algorithmes développés par nos collègues informaticiens et cogniticiens.