tribunes

Big data, data reuse en santé : un chemin semé d'embûches nécessitant une approche pluridisciplinaire

Illustration de la complexité d'utilisation de données de santé et de la nécessaire collaboration de plusieurs professions.

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Nous vous entraînerons avec nous dans un exemple de recherche ayant réussi. L'objet de cette recherche était de mettre en place un système d'intelligence artificielle (IA) permettant de prévenir les effets indésirables du médicament (EIM). Nous verrons que le chemin fut parsemé d'embûches. Ces embûches illustrent l'impossibilité de faire accomplir un tel travail par une machine seule, et l'absolue nécessité d'approches pluridisciplinaires.

Contexte et objectifs

En France, chaque hospitalisation de patient donne lieu à la collecte d'informations codées tels les diagnostics (par exemple « K37 » pour certaines appendicites) et les actes (par exemple « HHFA001 » pour certaines ablations chirurgicales de l'appendice). De plus, dans la plupart des hôpitaux, les médicaments sont prescrits *via* des logiciels et les résultats d'analyses de biologie médicale sont également transmis au service prescripteur par des logiciels spécifiques. Toutes ces données constituent des données massives ou *big data* (voir l'encadré de définitions) [3] : on dénombre en moyenne 100 résultats de biologie médicale par séjour (par exemple le taux d'hémoglobine) et 200 000 séjours par an dans certains CHU. Ces données servent respectivement à facturer le séjour à l'Assurance maladie et à soigner le patient. Elles

Quelques définitions simplifiées

Données massives/*big data*
Données volumineuses, comportant par exemple de nombreux individus et/ou de nombreuses variables, quelles que soient leur origine et leur exploitation.

Réutilisation de données/*data reuse*
Fait d'utiliser (pour de la recherche par exemple) des données qui ont initialement été collectées dans un autre but (le soin par exemple).

Intelligence artificielle (IA)
Fait de doter l'ordinateur de capacités visant à mimer le résultat d'une intelligence. Une IA peut s'appuyer sur un répertoire de règles écrites par un humain, un apprentissage automatisé (*machine learning*), ou un apprentissage par renforcement.

Effet indésirable du médicament (EIM)
Effet secondaire nocif et non voulu lié à la prise, la modification de dose ou l'arrêt d'un médicament (hormis erreurs d'administration et tentatives de suicide).

Emmanuel Chazard
Professeur des universités, faculté de médecine, Cerim ULR 2694, université de Lille, praticien hospitalier, CHU de Lille

peuvent cependant être réutilisées à des fins de recherche, on parle alors de réutilisation de données ou *data reuse*.

Notre objectif ici est de les réutiliser pour détecter automatiquement les circonstances causant des EIM. Une fois ces circonstances identifiées, nous bâtirons un logiciel d'IA [10] qui « surveillera » les prescriptions médicamenteuses et alertera le prescripteur en cas d'élévation du risque d'effet indésirable du médicament. Cet énoncé illustre le lien qu'entretiennent parfois *big data*, *data reuse* et intelligence artificielle. Cela dit, ce lien n'est pas obligatoire. La plupart du temps, le grand public parle de *big data* pour désigner le *data reuse* parce que les jeux de données dont la réutilisation est la plus prometteuse sont généralement de grande taille. Passées ces définitions, lançons-nous dans notre recherche !

Notre recherche, pas à pas

Construction d'un entrepôt de données

La première étape est « d'aspirer » les données disponibles et de les nettoyer (indifféremment de l'objectif). Copiées dans une nouvelle base de données plus simple, elles constituent un « entrepôt de données ». Première surprise, un certain nombre d'enregistrements sont « orphelins » et ne peuvent pas être rattachés à un séjour hospitalier… la complexité des systèmes d'information hospitaliers (SIH) est telle, qu'il est difficile de savoir si le lien existe mais est perdu lors de l'extraction, ou si l'information est d'emblée corrompue. Le chercheur, quant à lui, devra renoncer à tout analyser. Ensuite, les données comprennent de nombreuses erreurs. En voici un exemple : dans un hôpital, le libellé « hématies » est associé à des valeurs comprises entre 4 000 et 5 500/mm³. C'est mille fois trop faible, pourtant ces patients sont tous vivants. L'unité enregistrée est tout simplement fausse : pourtant, le médecin interprète le chiffre précisément mais sans regarder s'il s'agit de milliers ou millions, et parfois sans connaître précisément l'unité attendue. D'autres patients également ont un taux de potassium tournant aux alentours de 30 mmol/l, la norme devant rester inférieure à 4,5. Sont-ce des momies ? Non, simplement des patients dont on a dosé le taux de potassium urinaire, mais sans l'écrire clairement dans la base de données. Ces types d'erreurs n'ont généralement aucun impact sur le soin, car le soignant les corrige mentalement sans même s'en rendre compte. Elles sont critiques lorsqu'on analyse les données.

Leçon apprise

La construction d'un entrepôt de données requiert des informaticiens, mais aussi des médecins capables d'évaluer la qualité des données, et des analystes capables de dire quelles données sont indispensables ou peuvent être sacrifiées.

Analyse de données

Nos données sont enfin prêtes. Mais qu'en faire ? Certes, la construction d'un entrepôt de données nous a permis de passer de plus de mille tables de données d'un SIH à quelques dizaines de tables, mais les données restent complexes : hétérogènes (diagnostics, actes, médicaments, résultats d'analyses…), presque toujours manquantes et presque jamais par hasard (très peu de patients bénéficient d'une IRM de l'hypophyse… parce que leur hypophyse est normale), très complexes (près de 40 000 codes différents pour décrire les maladies, et encore cela ne suffit pas), et plus ou moins fiables !

Et pourtant, observons un médecin parcourir un dossier médical (figure 1) :

« Ce patient fait probablement une hémorragie : son **INR [International Normalized Ratio, l'un des indicateurs de la coagulation sanguine] augmente**, *témoignant d'une activité anticoagulante trop forte, et le lendemain son* **taux d'hémoglobine diminue**, *traduisant sans doute une hémorragie. Un* **antivitamine K** *(anticoagulant) était administré juste avant, et vraisemblablement potentialisé par du* **paracétamol à forte dose**. *Le médecin corrige ensuite cela en* **arrêtant l'anticoagulant** *et en* **introduisant de la vitamine K**, *son antidote. Par la suite, l'*INR **se normalise.** »

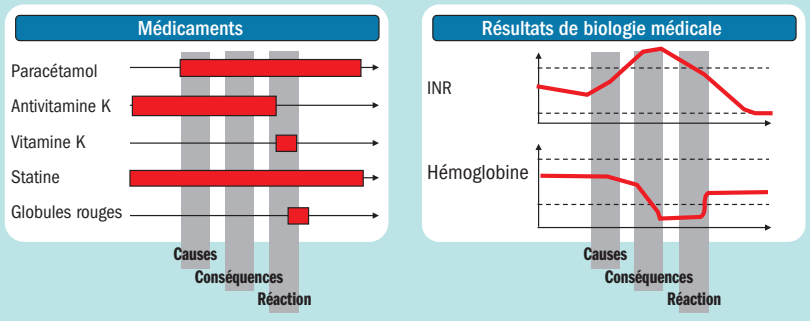
Quoi qu'on en dise, aucun ordinateur n'est capable de tenir un tel raisonnement. L'analyse de ce raisonnement nous permettra néanmoins de définir des « caractéristiques » (ou *features*) qu'il faudra calculer [13]. Il s'agit simplement d'appliquer tout un tas d'algorithmes (décidés par un expert) pour créer autant de variables simplifiées qu'il y avait de constatations **soulignées** ci-dessus, et plus encore par généralisation. Ces caractéristiques se présentent de manière plus simple que les données réelles disponibles, et elles sont « optimisées » pour porter du sens. Ainsi, grâce à l'extraction de caractéristiques, nos algorithmes de *machine learning* ne seront plus comme une poule face à un couteau.

Leçon apprise

L'extraction de caractéristiques nécessite en amont une excellente connaissance des données, une maîtrise de l'algorithmique de base, et en aval une connaissance des formes de données qui « fonctionnent » en *machine learning*.

figure 1

Interprétation de données brutes d'un patient qui présente une hémorragie sous anticoagulants



Lecture : le temps va de gauche à droite ; le commentaire est dans le texte

Prédiction par *machine learning*

L'extraction de caractéristiques nous a permis de construire un grand tableau, comprenant une seule ligne par patient, et des milliers de colonnes, représentant des variables simples. Nous avons supprimé la complexité des données. Il est à présent aisé d'utiliser des techniques d'apprentissage automatisé (*machine learning*) pour prédire automatiquement certaines variables (par exemple présenter une hémorragie) à l'aide de toutes les autres (avoir un anticoagulant, l'âge, le sexe, etc.). À ce jeu-là, les réseaux de neurones sont plutôt doués. Hélas, ils construisent pour ce faire des formules mathématiques de plusieurs pages, incompréhensibles par un humain. Cela leur vaut d'être qualifiés de « boîte noire ». En pratique, dans notre projet, ils s'avèrent inutilisables : comment alerter un médecin sur un risque, sans même être capable de livrer des arguments validés scientifiquement ? Nous préférons donc une technique moins vendeuse d'un point de vue marketing, mais plus lisible, comme les arbres de décision (figure 2). Les premiers arbres nous apprennent que… les patients plus âgés meurent davantage. Ce résultat était peut-être connu dès la Préhistoire ! Après des corrections de variables, les arbres suivants nous apprennent par exemple que le plus fort facteur de risque d'hyperglycémie, c'est d'avoir de l'insuline. Or l'insuline entraîne des hypoglycémies. C'est pourtant évident : association statistique ne signifie pas causalité. Il faut mettre en place des procédures automatisées et expertes de filtrage des associations découvertes par la machine. À la fin, nous tenons le bout (figure 2) : ainsi par exemple, sur sept patients présentant une insuffisance rénale et âgés de plus de 85 ans et ayant de la spironolactone, six (85 %) présentent ensuite une hyperkaliémie. L'association est techniquement, statistiquement et bibliographiquement valide. Pourtant, la relecture des dossiers nous apprend que seulement la moitié des pathologies sont réellement des effets imputables au médicament, car il arrive tout un tas d'autres choses à ces patients, invisibles dans les données.

Leçon apprise

La prédiction par *machine learning* n'est donc pas seulement un jeu de statisticien. Elle nécessite, outre l'extraction de caractéristiques, une forte expertise métier, des résultats analysables et critiquables par un humain, et une validation par retour aux cas réels.

Prévention en vie réelle

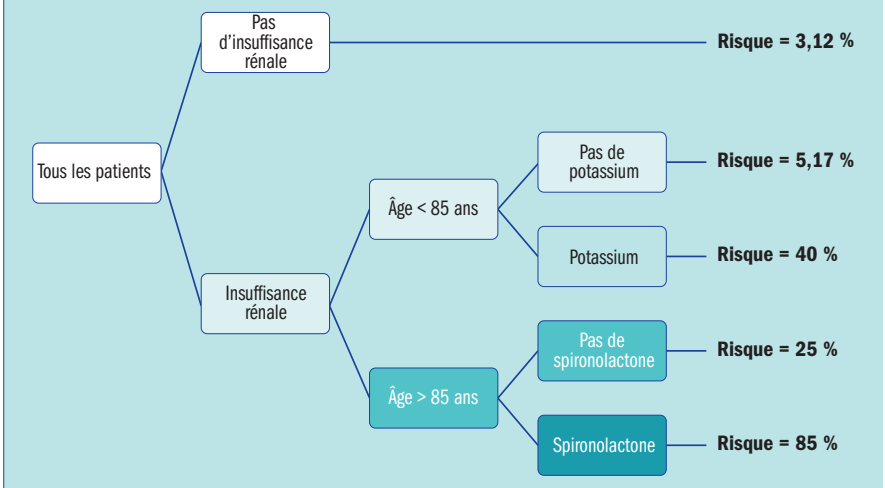
Au bout du processus de *machine learning*, nous tenons enfin un lot de 256 règles permettant de prédire la survenue d'EIM. Il « suffira » de les intégrer dans un système d'aide à la décision connecté au logiciel de prescription. Nous apprenons au passage que le niveau de risque (pourcentages dans la figure 2) varie fortement d'un service à l'autre. Plus généralement, les médecins maîtrisent très bien les médicaments de leur spécialité, et pour eux les alertes sont totalement inutiles : le fait est qu'elles leur pourrissent la vie sans diminuer les erreurs ! Inversement, c'est plutôt sur les situations plus rares qu'ils peuvent se tromper. Ainsi, par exemple, on ne verra jamais d'hyperkaliémie en néphrologie bien que tous les patients soient à risque, mais on peut en voir en pneumologie. Forts de ce constat, nous déployons le premier « SPC-CDSS » [12], c'est-à-dire le premier système d'aide à la décision filtré et contextualisé statistiquement. Bien que novateur et hautement valide, ce système ne sera pas utilisé par les cliniciens. Nous avions tout pris en compte, sauf le facteur humain. Bêtement, comme tant d'autres avant nous, nous avions engendré une monstruosité : un logiciel qui prend du temps à ceux qui en ont le moins, un logiciel qui rajoute des clics, des actions et des alertes à ceux auxquels tout le personnel hospitalier transfère insidieusement son travail : les médecins.

Leçon apprise

On ne peut pas se contenter de dire : « Le logiciel est bon donc ils l'utiliseront. » L'informatisation des hôpitaux a insidieusement transféré les tâches du personnel le moins qualifié vers le plus qualifié. Argumentant des économies, elle a augmenté le coût du travail pour une tâche donnée, et diminué la productivité pour certaines catégories de personnel hautement qualifié. Un bon logiciel d'IA est surtout un logiciel qui s'intégrera dans un *workflow* de manière à alléger la charge cognitive et l'agacement des professionnels qualifiés, et leur faire gagner du temps.

figure 2

Exemple d'arbre de décision (*machine learning*) prédisant automatiquement un risque d'hyperkaliémie



Conclusion

À travers ce cheminement, nous avons illustré les nombreux obstacles rencontrés, et la nécessité de faire collaborer au minimum informaticiens, statisticiens, spécialistes des facteurs humains, et surtout experts des données considérées (ici, des médecins). La vidéo citée en référence présente également ce cheminement [11].