Christian Lovis[1]

Données massives (Big Data) en santé

De quoi parle-t-on lorsqu'on parle de données massives, ou encore de «Big Data»? Depuis une décennie, la société s'informatise à un rythme soutenu, sinon effréné. Transactions financières, processus logistiques, administration en ligne, santé, tous les aspects de l'économie, du fonctionnement de la société et jusqu'à notre vie privée se digitalisent.

Et voilà qu'une des interrogations fondamentales de la pensée humaine prend une nouvelle dimension. Comme si bien exprimé par Pierre Simon Laplace dans son *Essai* philosophique *sur les probabilités*: «Une intelligence qui, à un instant donné, connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était suffisamment vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome; rien ne serait incertain pour elle, et l'avenir, comme le passé, seraient présents à ses yeux» [1]. C'est le vieux rêve, ou peut-être le cauchemar, du déterminisme, de la capacité à comprendre, mais aussi à prédire, qui s'en trouve bouleversé.

> Les espoirs sont immenses, à juste titre, mais les défis sont à la hauteur des espoirs.

En santé, cette évolution s'accompagne de perspectives inédites, qui se cristallisent autour du concept de «médecine de précision», ou encore «médecine personnalisée». C'est la convergence de toutes les données qui caractérisent un individu, de ses gènes à son corps, son environnement, son style de vie, l'écosystème dans lequel il vit, aux fins de pouvoir au mieux et de manière absolument individualisée comprendre ses besoins, soutenir les démarches médicales et thérapeutiques, interpréter les résultats, mais également, et peut-être surtout, maintenir son capital santé et entrer dans une démarche de prévention de la maladie.

Il y a un nombre croissant d'exemples des bénéfices de l'utilisation de données massives pour la santé comme en pharmacogénétique [2] qui renforcent l'espoir de pouvoir ainsi significativement soutenir et améliorer la santé, l'efficience et l'efficacité du système de santé en général.

Toutefois, la capacité à pouvoir exploiter ces données pour en tirer de la connaissance, tant à l'échelle de l'individu que pour une population ou un système, dépend de la capacité à analyser, interpréter et utiliser ces données et d'intégrer de nombreuses sources souvent disparates et hétérogènes, de qualité variable et souvent indéterminable, qui couvrent un immense spectre allant de la molécule à l'écosystème.

Le terme de «Big Data» est né en 2001 avec la définition des trois «V» proposée par Doug Laney [3]: «Volume, Vitesse, Variété». Dans cette vision, Doug Laney considère les aspects suivants:

1. (V)olume, car il s'agit de très grands volumes de données;
2. (V)itesse, car il convient de pouvoir analyser et traiter ces données en un temps raisonnable;
3. (V)ariété, car elles sont disparates et hétérogènes.

Plus tard, les 3 «V» ont été bien étendus, et notamment la Vitesse se réfère aussi à la notion de données en flux, par exemple des données produites de manière continue. Aujourd'hui, certains auteurs évoquent jusqu'à 11 «V», dont – entre autres: (V)aleur, (V)ariabilité et (V)éracité.

Les défis

Les espoirs sont immenses, à juste titre, mais les défis sont à la hauteur des espoirs. Ces défis se résument en un mot: interprétabilité. Afin d'illustrer le propos, imaginons que d'un coup de baguette magique il soit possible d'avoir une copie de tous les résultats d'analyse de laboratoire pratiqués en Suisse depuis 10 ans sur un seul serveur. A première vue, on est tenté de penser que ce serait génial, mais bien vite se poserait la question: qu'en faire? Comment les interpréter? Comment les agréger? Différentes manières de nommer les analyses, différentes méthodes, unités et valeurs normales, contextes variables voire inconnus, raisons ayant motivé les analyses, situation des patients, etc. entre autres!

Les défis des données massives sont, pour part, celles qui sont habituelles en biostatistiques. Toutefois, les données massives soulèvent aussi des questions nouvelles. Voici une liste de quelques défis nouveaux:

Protection de la sphère privée

Une des caractéristiques intéressantes de l'approche en données massives est la capacité de connecter des données issues de multiples sources. Par exemple, connecter les déplacements venant des téléphones cellulaires, des achats cumulés pour son alimentation et le dossier médical pour identifier de nouveaux facteurs de risque et d'exposition. Cette capacité à connecter le «data linkage» est vrai pour l'individu, mais aussi sa famille, des groupes d'individus, etc.

Qualité

La qualité des données est un problème bien connu et récurrent [5] qui

1 Prof. Christian Lovis, Service des sciences de l'information médicale, Hôpitaux universitaires de Genève

soulève un tout nouvel aspect. L'utilisation des réseaux sociaux, par exemple des tweets, a montré sa capacité à identifier des comportements liés à la santé [6]. Il est impossible de modifier ou d'influencer la qualité de ces données; les personnes qui écrivent sur des réseaux utilisent leur manière de s'exprimer, et il faut faire avec. Même pour des données fiabilisées, comme les prescriptions médicamenteuses, la qualité de ces informations variant avec de nombreux facteurs dans le temps, l'utilisation de systèmes informatisés, les alertes, les compétences des utilisateurs, etc. Dès lors, le défi devient plutôt d'être capable de décrire la qualité de chaque source de données et son évolution dans le temps, plutôt que de s'attacher à augmenter la qualité de ces informations.

Analytique

L'analyse de données massives doit également affronter de nouveaux défis qui vont bien au-delà des défis traditionnels rencontrés en sciences de la vie. Il va falloir développer des méthodes robustes et fiables d'analyse distribuée, domaine très peu exploré à ce jour. Ensuite, il va falloir mettre en place des méthodes analytiques incrémentales, surtout lorsque les sources sont en flux. Il sera rapidement impossible de traiter les volumes requis constamment et en temps réel, il faudra donc pouvoir faire de petits traitements successifs qui «s'ajoutent» les uns aux autres [7, 8]. Ici aussi, il s'agit d'un domaine qui commence juste à se développer. Finalement, de nouvelles méthodes d'apprentissage sont requises en regard de la masse d'informations à disposition, comme par exemple le «*deep learning*» [9].

Sémantique et interopérabilité

La maîtrise de la sémantique est une condition sine qua non à l'interopérabilité. Toutefois, cette maîtrise est loin d'être simple, ni d'ailleurs acquise. L'interopérabilité reste le défi mondial numéro 1 dans l'informatisation en santé. Depuis 30 ans, de nombreuses initiatives (HL7, IHE, Meaningful Use, EPSOS, EXPAND, e-health-suisse.ch), publiques comme privées, ont tenté de résoudre ce problème. Le fait demeure que cela reste un réel obstacle et que les progrès sont lents, car ils requièrent la convergence de moyens techniques, de standards sémantiques telles des ontologies, et d'environnements régulatoires et notamment incitatifs.

Données textuelles

La plus grande partie de l'information réside dans ce qui est généralement considéré comme non structural, et en particulier les textes. De fait, la manière dont on considère cette information est en passe de changer radicalement, et le texte est de plus en plus considéré comme étant une source primaire pour l'analyse. Ceci est surtout lié à l'amélioration des performances de calculs des machines et des outils de traitement automatique du langage naturel ou des images [11, 12]. Récemment, le système IBM Watson a démontré la puissance de cette approche en battant des humains à un jeu télévisé aux USA sur une base essentiellement d'apprentissages probabilistiques de textes provenant de sources accessibles en ligne [13].

Conclusions

L'ère de la «data driven science» s'ouvre, avec de fantastiques perspectives pour améliorer notre compréhension de la vie, de son fonctionnement, mais aussi d'appliquer et de mesurer ces connaissances pour les individus et les populations. Toutefois, comme brièvement évoqué, les défis à relever sont à la hauteur des espoirs. Une des conséquences de l'usage de données massives est également qu'il sera de plus en plus difficile, voire impossible, de vérifier les résultats par des études prospectives. Il faudra donc aussi repenser les méthodes de validation, les responsabilités et l'usage des résultats. Il est donc important de soutenir des programmes de recherche dans ce domaine, mais aussi de mettre en place des mesures incitatives fortes pour former à ces nouvelles approches et construire les compétences requises.

Correspondance:
Christian.Lovis@hcuge.ch

Références

Vous trouverez la liste des références sur le site: www.sulm.ch/f/pipette → Numéro actuel (n° 2-2016).

Datenflut im Gesundheitswesen (Big Data)

Das Zeitalter der «datengetriebenen Wissenschaft» bricht an und eröffnet grossartige Perspektiven für ein besseres Verständnis des Lebens und seiner Funktionsweise, aber auch die Anwendung und Bewertung dieser Erkenntnisse für den einzelnen Menschen und das Gemeinwesen. Die Aufgaben, die bewältigt werden müssen, sind jedoch ebenso gross wie die Hoffnungen. Eine Folge der Verwendung von Massendaten ist ausserdem, dass es zunehmend schwierig, wenn nicht gar unmöglich sein wird, Ergebnisse durch prospektive Studien zu überprüfen. Auch die Validierungsmethoden, die Verantwortlichkeiten und die Verwendung der Ergebnisse müssen demnach überdacht werden. Diesbezügliche Forschungsprogramme müssen gefördert, aber auch starke Anreize geschaffen werden, diese neuen Ansätze und Methoden zu erlernen und die nötigen Fähigkeiten zu erwerben.

Focus Swiss MedLab 2016: BIG DATA

Keynote-Referat:
«BIG DATA im Gesundheitswesen: Hoffnung und Herausforderung»
Prof. Christian Lovis

Parallel Sessions:
Nutzen von Biobanken für die moderne Labordiagnostik, Prof. Vincent Mooser, Prof. Thomas Illig, Prof. Michael Hummel
Vermessung des Ich und die Rolle der Labordiagnostik in der Zukunft, Peter Ohnemus, Prof. Michael Lehmann, Catherine Bugmann
Big Data in Genomik und Krebsbehandlung
Prof. Torsten Haferlach, Prof. Aurel Perren, Dr. Ursula Amstutz
Companion Diagnostics
Dr. Gabriele Beer, Prof. Peter Johannes Wild, Dr. Axel Nemetz

Datum: Mittwoch, 15.6. & Donnerstag, 16.6.2016
Sprache: Deutsch, Französisch, z.T. Simultanübersetzung

Weitere Infos:
www.sulm.ch/swissmedlab