

Big Data dans la santé, une quadruple révolution : scientifique, médicale, technologique et économique

Yann Gaston-Mathé, HyperCube Research (yann.gaston.mathe@gmail.com)
Patrice Denéfle, Ipsen (pdenefle@gmail.com)

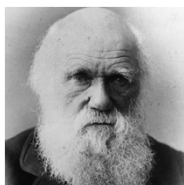
Colloque XMP-Biotech,

17 mars 2014

Les tendances à l'œuvre :

économiques, technologiques, scientifiques, culturelles et ...

... humaines



Population height difference between
North and South Korea cannot be
explained by genetic evolution !

Une époque charnière pour le monde la santé

Tendance économique :

- Explosion des coûts de la santé
- Pression de plus en plus forte sur les prix, les coûts, les remboursements
- Vers un système de plus en plus « dirigiste »

Tendance « techno-sociologique » :

- Informatisation, dossiers électroniques, interopérabilité
- Multiplication des données personnelles
- E-health, mobile solutions
- Vers le partage et l'intégration

Vers la
médecine
personnalisée

Tendance « techno-scientifique » :

- Technologies « omiques »
- Vers le génome à 1000\$
- Vers la généralisation de l'utilisation des données génétiques dans la pratique clinique

Le Big Data dans la santé est une tendance de fond

A radical change in Medical practice

Leroy Hood : the P4 Medicine (<http://p4mi.org/>):
Predictive, Preventive, Personalized, Participatory

Systems Biology and Systems Medicine
Consumer-Driven Healthcare and Social Networks
Digital Revolution
P4 Medicine
Change agent

http://www.ted.com/talks/dave_debronnkart_meet_e_patient_dave

J Med Internet Res. 2008 Jul-Sep; 10(3): e22.
Published online Aug 25, 2008. doi: 10.2196/jmir.1030

Big data dans la santé, quel progrès médical en attendre ? 4 Colloque XMP Biotech, 17 mars 2014

La transformation de la santé sera portée par l'amélioration de la compréhension des phénomènes à 3 niveaux

Clinique :
Patient
Maladie Médicament

Biologie :
Maladie
Environnement Gènes

Système de santé :
Bénéfice
Risque Coût

Big data dans la santé, quel progrès médical en attendre ? 5 Colloque XMP Biotech, 17 mars 2014

La révolution de la santé passera par une intégration de la chaîne de soins, et donc une consolidation des systèmes d'information

Scattered patient care landscape

Integrated Patient care value chain

Early detection of pre-symptoms
Disease onset prevention and detection
Personalized diagnostic
Prognostication and treatment selection
Therapy (first line)
Chronic care
Post-treatment monitoring
Dependency care

Integrated holistic patient / disease management

La consolidation des données devient un enjeu majeur

Ce que cela suppose :

- En amont (recherche), la production de connaissances à partir des données consolidées et mutualisées
- En amont : la mise à jour et la diffusion permanente des connaissances
- En aval : la maîtrise des flux de données et la coordination effective des acteurs du système de santé

Big data dans la santé, quel progrès médical en attendre ? 6 Colloque XMP Biotech, 17 mars 2014

Mieux comprendre la réalité clinique et biologique : une nécessité pour optimiser la balance bénéfice-risque-coût

La réalité clinique et biologique est complexe...

- Hétérogénéité des situations et des populations
- Multi-causalité des phénomènes : génétiques, environnementales, individuelles, collectives...

... et elle est minorée par les essais cliniques...

- Qui sélectionnent de petites populations homogènes dans un cadre très contrôlé
- Dont les analyses « post-hoc » sont souvent limitées

... d'où des écarts importants fréquemment observés dans la pratique clinique...

- Certains médicaments « déçoivent »
- Certains médicaments « marchent mieux »

... et la remise en cause par les Autorités ou les payeurs de la balance bénéfice-risque-coût

Questions clés:

- Quelles sont les combinaisons de facteurs qui favoriseraient une efficacité optimale du médicament ?
- Y-a-t-il des sous-populations qui répondent mieux (moins bien) ?
- Y-a-t-il des associations de médicaments/traitements défavorables à l'efficacité ?
- Dans le cas de survenue d'événements indésirables préoccupants, quels sont les profils des patients concernés ?
- Quel est l'impact économique du traitement ?

Les laboratoires pharma doivent apporter des réponses à des questions clés pour lesquelles l'accès aux « real world data » devient indispensable

Big data dans la santé, quel progrès médical en attendre ?

7

Colloque XMP Biotech, 17 mars 2014

La croissance sans fin du volume et de la variété des données de biologie moléculaire ouvre la voie à des progrès majeurs...

Les données de la biologie moléculaire

Le génome : la séquence de l'ADN

- Génotypage: SNP, cGH-Array, CNV,...

• Séquençage:

- Génome complet (2,3 milliards de bases) (\$1000 genome)
- Exome (les parties codantes de l'ADN = 1% du total), RNAseq, ...
- Ciblés sur des gènes candidats

Le transcriptome : quantification des ARN messagers (23.000 gènes, >100,000mRNA), rétention d'intron, splicing alternatif, ...

Le méthylome : analyse de la méthylation de l'ADN et des histones, ...

Les **micro-RNAs**, les **exosomes**, ...

Le protéome : analyse, caractérisation, quantification des protéines de l'organisme

Le métabolome : les métabolites de l'organismes

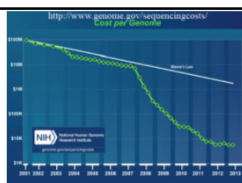
Le microbiome : la flore bactérienne

Le répertoire immunitaire

Le connectome : l'analyse des connexions neuronales

Etc...

... mais au prix d'une complexité extraordinaire



Big data dans la santé, quel progrès médical en attendre ?

8

Colloque XMP Biotech, 17 mars 2014

The advent of genomic technologies has driven a profound redefinition of oncology taxonomy

- **Initial cancer taxonomy (1900-1960s):** based on **organ**, **cell type**, **tissue invasion** (metastases) and **histology** (tumor cell differentiation)
- **40 years ago: Sakurai and Avery (1973):** Specific tumor changes (**karyotype**) could have an impact on clinical management of leukemia patients
- **Bladder cancer**, first solid tumor for which **genomic abnormalities** were incorporated in the clinical prognostic scheme (Falor et al, 1978)
- The **genome theory of cancer** has gained support over the next 20 years (1980-2000) because of the **persistent failure** of cytotoxic chemotherapies
- The revolution occurs because of the Consilience between:

Tumor biology

Genome technology

Computational innovation

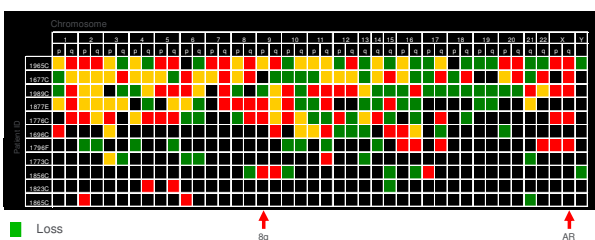
drug discovery

Big data dans la santé, quel progrès médical en attendre ?

9

Colloque XMP Biotech, 17 mars 2014

Example: Advanced disease patients have multiple cancers/genomic changes
(R.Vesella et al.)



➡ Gains at 8q and androgen receptor (AR) region is associated with hormone refractory prostate cancer ¹.
➡ Which alterations are specifically related to progression?

Big data dans la santé, quel progrès médical en attendre ?

10

¹Napponen et al., Am. J. of Pathol., 1998.

Colloque XMP Biotech, 17 mars 2014

Aux US, l'émergence du business des données patients et du « Healthcare big data analytics »

Healthcare providers

- Geisinger
- Kaiser Permanente

Health Insurers

- Aetna – Aetna Innovation Labs

State Health Insurance

- Idaho Health Data Exchange

Service providers

- 23andMe

Integration of Patient Generated Health Data (PGHD)

- E-health initiatives

Federal State

- White House Big data R&D initiative

Reduce RISKS

Improve HealthCare

Control Costs

Big data dans la santé, quel progrès médical en attendre ?

11

Colloque XMP Biotech, 17 mars 2014

Quelles perspectives de progrès pour la médecine ?

Les perspectives ouvertes par le Big data dans la santé :

1. Mieux comprendre

- les mécanismes fondamentaux des maladies, les interactions entre facteurs génétiques, épigénétiques, environnementaux, pour permettre :
 - o la poursuite de l'innovation thérapeutique
 - o la prévention et l'élimination des facteurs de risques
- Les mécanismes de toxicité médicamenteuse, etc.

2. Prédire les phénomènes, afin :

- d'anticiper la survenue de problèmes de santé,
- d'administrer le bon traitement au patient, etc.

3. Alerter en temps réel, sur :

- la dégradation de l'état de santé afin d'accélérer la prise en charge des patients,
- la survenue d'épidémies
- ...

Big data dans la santé, quel progrès médical en attendre ?

13

Colloque XMP Biotech, 17 mars 2014

1. Mieux comprendre - Exemple d'application :

Paving the way to the understanding and treatment of complex diseases: dynamic simulation

Vol 14(12) April 2014 | DOI:10.1002/hsb.20008

FUTURE

ARTICLES

Direct analysis of the genome sequence differences (500 Gbp) between

- primary,
- xenografted
- and eventually metastatic tumours



Understanding of the natural history of individual patients tumor

Genome remodelling in a basal-like breast cancer metastasis and xenograft

Li Ding^{1,2*}, Matthew J. Ellis^{1,2*}, Shuangge Gao¹, David E. Lanoie¹, Ken Chen¹, John W. Wylie^{1,3}, Christopher C. Harris⁴, Michael D. McLellan¹, Robert S. Fulton^{1,2}, Lucinda L. Fulton^{1,2}, Rachel M. Abbott¹, Jeremy Heng¹, David J. Douville^{1,2}, Daniel C. Kobayashi¹, Heather Schmidt¹, Justin Sakash¹, Congyan Zhang¹, Lin Chen¹, Ling Lin¹, Matthew C. Weng^{1,2}, Joshua F. McMichael¹, Vincent J. Maguire¹, Lisa Cook¹, Sam D. McClellan¹, Tyrone L. Vignery¹, Elizabeth Applebaum¹, Katherine C. Brown¹, Shou Chen¹, Thomas C. Chen¹, Li-Lian Raker-Cowder¹, Yu Tao¹, Benjamin E. Jager¹, Scott M. Smith¹, Adam T. Coker¹, Robert E. Johnson¹, Craig E. Pohl¹, Eric D. Candeskey¹, Carina C. Frankel¹, Kimberley A. Page¹, Jerry S. Bonf¹, Judy S. Robinson¹, Jennifer S. Hodges¹, William Schorling¹, Nathan D. Chao¹, Dong Shou¹, David F. Locke¹, Madeline E. Watson¹, James M. Eide¹, Jack B. Peck¹, Benjamin J. Chertkov¹, Justin T. Lofgren¹, Felix Du¹, Amy E. Hawkins¹, Michelle D. O'Laughlin¹, Kelly E. Bernard¹, Mark Cunningham¹, Glendora Elbert¹, Mark D. Moore¹, Christine M. Thompson¹, Jennifer L. Lian^{1,2,3,4}, Paul J. Goodfellow¹, Charles M. Perou¹, George M. Vlastakis^{1,2}, Rebecca Ait¹, Mark Watson¹, Timothy J. Ley^{1,2,3,4}, Richard A. Wilson^{1,2,3,4}, & Elaine R. Mardis^{1,2,3,4}

Massively parallel DNA sequencing technology provides an unprecedented ability to screen entire genomes for genetic changes associated with tumor progression. Here we describe the genomic analysis of four DNA samples from an African-American patient with basal-like breast cancer: peripheral blood, the primary tumor, a brain metastasis and a xenograft derived from the primary tumor. The metastases contained two to six new mutations and a large deletion not present in the primary tumor, and was significantly enriched for 20 genetic mutations. The xenograft retained all primary tumor mutations and displayed a mutation enrichment pattern that resembled the metastasis. Two overlapping large deletions, encompassing CDKN1A, were present in all three tumor samples. The differential mutation frequencies and structural variation patterns in metastasis and xenograft compared with the primary tumor indicate that secondary tumors may arise from a diversity of cells within the primary tumor.

Big data dans la santé, quel progrès médical en attendre ?

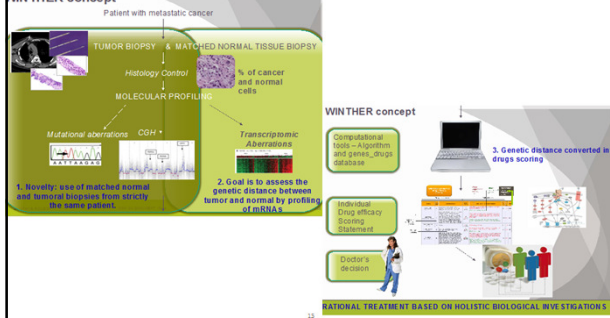
14

Colloque XMP Biotech, 17 mars 2014

1. Mieux comprendre - Exemple d'application :

Essai WINTHER : Choisir le traitement le plus adapté au profil "moléculaire" de la tumeur

WINTHER concept



Big data dans la santé, quel progrès médical en attendre ?

15

Colloque XMP Biotech, 17 mars 2014

2 – Prédire pour optimiser la prise en charge - Exemple d'application :

Computer-assisted Medical diagnostic & therapeutic algorithm recommendation

Watson is a natural language processing supercomputer (2880 Cores, Synapse)

Test phase 0: Jeopardy TV-contest

Test Phase 1:

- Sloan Kettering & MD-Anderson cancer centers.
- Watson, a Medical student has passed qualification as a assistant professor
- Watson can cross 20,000 cancers cases or 200M medical record pages in 15 sec

Test Phase 2:

- Call for innovative programs
- Financial Business analytics
- Watson 2.0: image processing

Next?

- Watson 3.0: Cloud and tablet based

IBM's Watson uses Jeopardy skills to become House-like medical diagnostician



<http://www.engadget.com/2013/10/15/ibm-watson-medical-diagnosis/>

Big data dans la santé, quel progrès médical en attendra-t-on ?

16

Colloque KMP Biotech, 17 mars 2014

2 – Prédire pour optimiser la prise en charge - Exemple d'application :

Un nouveau test de prédisposition génétique à l'autisme comprend plus de 1700 SNPs



IntegraGen Announces Launch of the ARISK2® Test

New version of test expanded to include children who have signs of developmental delay or ASD

IntegraGen has announced the availability of a new version of its ARISK2 Test that enables caregivers to assess the risk of ASD in children who have signs of developmental delay or ASD. The ARISK2 Test comprises over 1,700 genetic markers (single nucleotide polymorphisms, or SNPs) which are

analysed in order to identify a child's genetic risk score for ASD. DNA samples for the test are collected from children via a buccal swab.

A CLIA-certified laboratory genotypes DNA samples for the ARISK2 Test using a customized Illumina Infinium BeadChip.



To request ARISK2 Test DNA collection kits for your practice, call (877) 982-7475 Ext. 3 or order online at www.arisktest.com

L'autisme touche plus d'un enfant sur 100 et sa prévalence a presque doublé aux US dans les 30 dernières années.

L'autisme est une maladie multi-factorielle, complexe, mais la composante génétique est très importante : dans les familles avec un enfant autiste, le risque d'autisme pour les frères et sœurs est multiplié par 10.

Les thérapies comportementales ont un impact majeur sur l'évolution de la maladie et le niveau de handicap quand la prise en charge est réalisée précocement.

D'où l'intérêt de disposer de tests performants pour identifier les enfants à risque.

[http://www.arisktest.com/upload/pdf/ARISK20NewsLetter%20\(Vol%20No1\).pdf](http://www.arisktest.com/upload/pdf/ARISK20NewsLetter%20(Vol%20No1).pdf)

Big data dans la santé, quel progrès médical en attendra-t-on ?

Le test Arisk2 de la société IntegraGen permet de quantifier le risque de développer la maladie chez de jeunes enfants qui présentent un retard de développement ou des symptômes associés à l'autisme.

Le test Arisk2 est construit sur l'étude des polymorphismes génétiques (mutations ponctuelles : SNPs) fréquents chez les autistes vs les individus sains.

Comprenant 1700 SNPs qui forment un « score de risque », le test Arisk2 permet de bien discriminer les enfants autistes des individus sains ... sans pour autant accéder à la compréhension des mécanismes biologiques de la maladie.

17

Colloque KMP Biotech, 17 mars 2014

3. Alerter en temps réel - Exemple d'application

L'analyse en temps réel des requêtes Google permet de prédire de manière précise les épidémies de grippe saisonnière

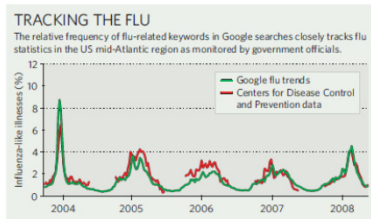
L'analyse des requêtes Google sur une liste de mots clé pertinents permet de détecter l'apparition des épidémies de grippe, avec 1 à 2 semaines d'avance sur le CDC (Center of Disease Control and Prevention).

Avantages :

- Vitesse
- Faible coût
- Précision géographique

Inconvénients :

- Risque de "faux positifs"



Source: <http://www.nature.com/news/2008/081119/pdf/456287a.pdf>

Big data dans la santé, quel progrès médical en attendra-t-on ?

18

Colloque KMP Biotech, 17 mars 2014

3. Alerter en temps réel - Exemple d'application

Les pacemakers connectés alertent en temps réel sur la dégradation du muscle cardiaque

Accent Pacemaker: simplified Patient Management From Implant Through Follow-up

"The Accent family of low voltage devices delivers simplified patient management from implant through follow-up."

Accent RF pacemaker features daily wireless remote monitoring, providing timely notification of actionable events and flexible remote follow-up scheduling.

Accent pacemaker (inductive) is the industry's smallest dual chamber and single chamber device capable of remote follow-up through Merlin.net™ Patient Care Network (PCN)."

FDA approved in 2009.

Competitors Boston Scientific and Medtronic have launched similar products

iPad Apps to monitor remotely pacemakers have been developed.



<http://professional.sjm.com/products/cm/pacemakers/dual-and-single-chamber/accel>
<http://www.fox.com/insider/general/2014/03/13/how-the-internet-of-things-will-connect-our-bodies.aspx>

Big data dans la santé, quel progrès médical en attendra-t-on ?

19

Colloque XMP Biotech, 17 mars 2014

3. Alerter en temps réel - Exemple d'application

Diabeo, application mobile pour diabétiques, permet un suivi en temps réel des paramètres et aide le patient à optimiser en temps réel et dans la durée le contrôle de sa maladie



Big data dans la santé, quel progrès médical en attendra-t-on ?

20

Colloque XMP Biotech, 17 mars 2014

Quels impacts pour la Recherche & Développement pharmaceutique ?

Data is essential to augment Information based decisions

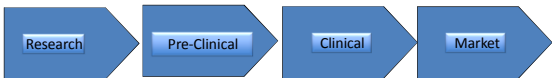
**The Goal:
Informed Decisions**



- Right Target
- Right Compound
- Right Disease
- Right Dose/Formulation
- Right Trial Design
- Right Patient Segment

Rig data dans la santé, quel progrès médical en attendre ? Colloque XMP Biotech, 17 mai 2014

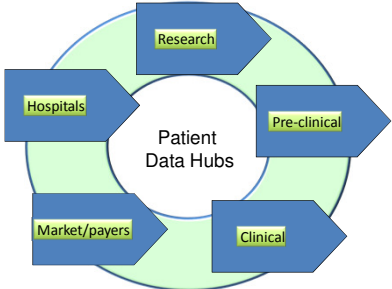
From the traditional Pharma R&D model...



Anticipation of true medical needs? Learning from Clinical experience?

23

... To a patient/data centric model



24

Boosting innovation by leveraging existing data: Computer assisted target discovery

- 23 Million scientific papers in Medline
- 1 Million papers published every year
- Text mining and prediction on 60'000 papers enabled to identify 10 potential kinases connected to p53
- 7/10 predictions turned out to be correct

MIT Technology Review

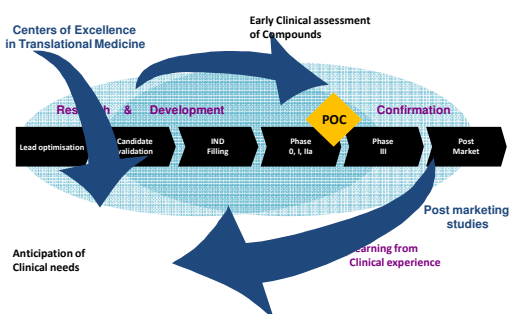
Software Mines Science Papers to Make New Discoveries

Software digests thousands of research papers to accurately identify proteins that could be productive targets for cancer drugs.

By Tom Simonite on November 25, 2013

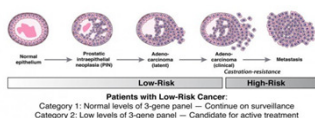
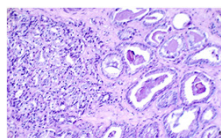
Software that read tens of thousands of research papers and then predicted new discoveries about the workings of a protein that's key to cancer could herald a faster approach to developing new drugs.

Digital & Translational Sciences Network interactions at all stages of R&D

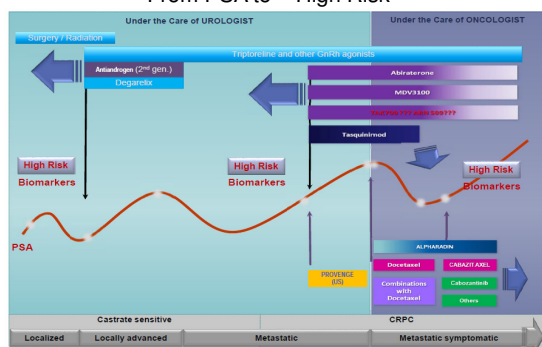


26

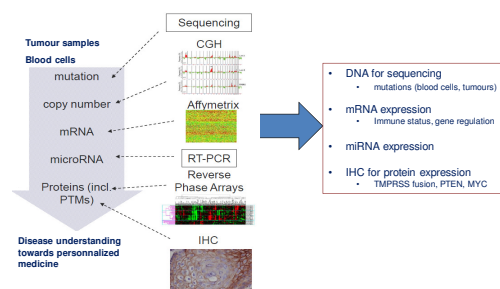
Prostate Cancer Case study



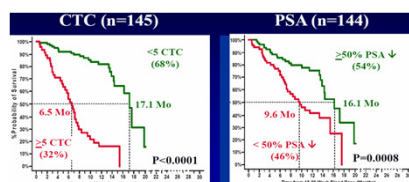
Changes in Prostate Cancer treatment landscape: From PSA to « High Risk »



Biological relapse Signal detection

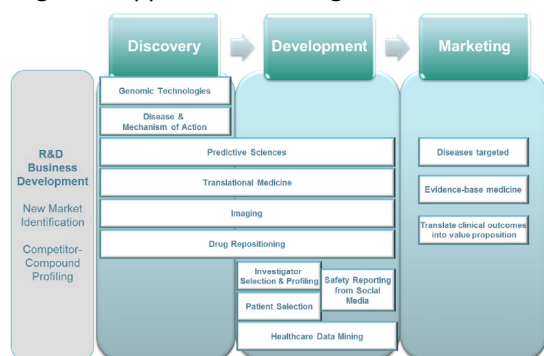


Example: CTC count as a Progressive Disease marker in Prostate cancer



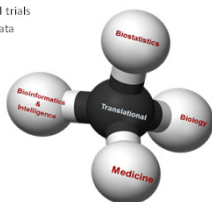
CTC number post treatment is more predictive of survival than a >50% decline of PSA from baseline

Big Data Opportunities along the value chain



Impact

- Late stage clinical trial outcome analysis**
 - Signal detection analysis
 - Meaningful insight into market value translating into future business
- Early stage Clinical trial design**
 - More accurate modeling and simulation of clinical trials
 - Contextualisation of internal data with external data
 - Value proposition assessment
- New R&D Project/target selection**
 - Early Safety target assessment
 - Translational Research plan
 - Medical need assessment



32

Quels enjeux, quels challenges ?

How Big is Big Data?

**Will Computers Crash Genomics?**

New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging.

11 FEBRUARY 2011 VOL 331 SCIENCE www.sciencemag.org

Big data dans la santé, quel progrès médical en attendre ?

34

Colloque XMP Biotech, 17 mars 2014

Des enjeux majeurs en terme d'informatique

De nouvelles données

- De toutes natures
- De tous formats
- Multi-sources

Les enjeux

- Accéder aux données
- Savoir les gérer
- Savoir les analyser pour en tirer de la valeur

→ Les verrous à lever :

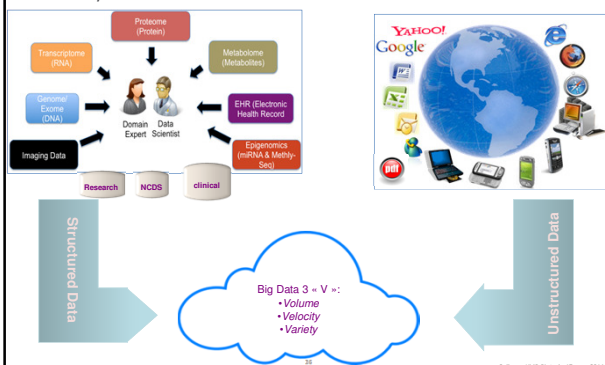
- Technologiques : puissance de calcul, stockage, parallélisation (Hadoop, NoSql, MapReduce, etc.)
- Scientifiques : algorithmes de data mining et machine learning
- Normatifs : standardisation des formats, interopérabilité, standards de qualité, meta-data
- Juridiques : propriété des données (IP), protection des données personnelles
- Gouvernance, humains et organisationnels
- Culturels : évolution inéluctable de nos systèmes de santé

Big data dans la santé, quel progrès médical en attendre ?

35

Colloque XMP Biotech, 17 mars 2014

An unprecedented volume and variety of data to store, consolidate and analyze...

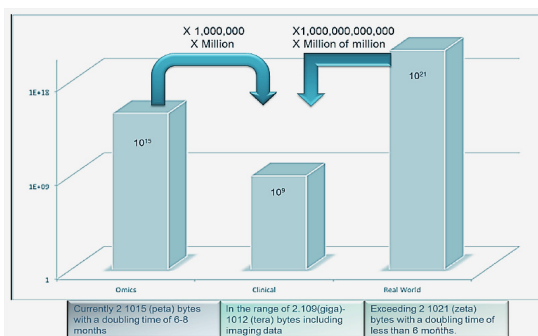


Big data dans la santé, quel progrès médical en attendre ?

36

Colloque XMP Biotech, 17 mars 2014

Des enjeux de stockage et de consolidation...



Big data dans la santé, quel progrès médical en attendre ?

17

Colloque XMP Biotech, 17 mars 2014

La "libération des données" et l'open innovation, des enjeux majeurs

The Sage BioNetworks initiative

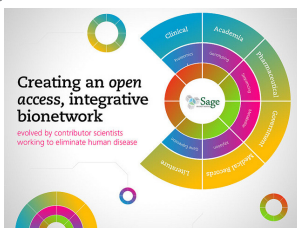
Created in 2009 as a non-profit organisation, Sage BioNetworks aim at redefining biomedical research through open systems, incentives, and norms.

Key obstacles to the emergence of "open source biology" according to Sage promoters:

- Technical, like file formats and ontologies.
- Career-based, like worries about getting scooped, or having someone else find the gold in your work.
- Legal, ranging from informed consent to trade secret.
- Lack of computable health data availability

Sage BioNetworks works:

- Research and open access publications
- Synapse technology platform, a collaborative set of technical services that allow scientists to access, share and analyze data together. 10'000 data sets available.
- Computational biology challenges
- Development of policies and procedures around the intake of new data and its publication for reuse.



Big data dans la santé, quel progrès médical en attendre ?

18

Colloque XMP Biotech, 17 mars 2014

Can we trust the data and the results from publications?

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Estimate: >50% of results published in scientific papers cannot be reproduced

Reasons:

- Fraud (minor issue)
- Incentives for scientists to publish their work
- Positive results have more impact than negative results
- Flaws in statistical methodology of experimental design
- Restriction in the size of Materials and methods sections in scientific journals
- Trend to generalize conclusions beyond the context of data generation
- ...



The Reproducibility initiative launched in 2013 aims at encouraging systematic independent validation of scientific results by other teams.

The American NIH has launched several initiatives (e.g., DDI, Data Development Index) to address this concern.

<http://www.reproducibility.org/>

<http://www.nih.gov/research-recovery/reproducibility-initiative/>

<http://www.nih.gov/research-recovery/reproducibility-initiative/>

<http://www.nih.gov/research-recovery/reproducibility-initiative/>

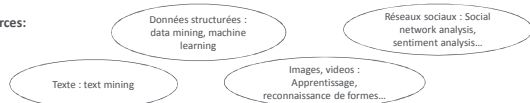
Big data dans la santé, quel progrès médical en attendre ?

19

Colloque XMP Biotech, 17 mars 2014

Technologies d'analyse de données: maîtriser la complexité des phénomènes est l'enjeu de la recherche de demain

Les sources:



Les contraintes :

- Hétérogénéité des populations et des causes des phénomènes
- Phénomènes liés à des interactions
- Phénomènes probabilistes
- Beaucoup de variables, peu d'observations (high dimensionality data)
- Des données de qualité hétérogènes, des données manquantes, du bruit
- Des données de type et de natures différentes

Difficultés et enjeux :

- Volumétrie, puissance de calcul nécessaire
- Détection des « **signaux faibles** », surtout les effets résultants d'interactions entre variables
- Capacité à prédire
- Risque de « **fausse découverte** » : capacité à généraliser les résultats (surapprentissage)
- Interprétabilité des résultats

Big data dans la santé, quel progrès médical en attendre ?

40

Colloque XMP Biotech, 17 mars 2014

L'Analytique, un domaine de recherche très fécond porté par le Web et la Bioinformatique

3 grandes familles

L'apprentissage non supervisé (clustering)

Les méthodes de classification / prédiction :

- Méthodes fondées sur la régression statistique
- Réseaux de neurones
- Méthodes bayésiennes
- Arbres de décision
- Méthodes ensemblistes (Random forests...)
- Support Vector Machines

L'extraction de règles (supervisées ou non)

- Règles non supervisées : Frequent ItemSets
- Règles supervisées : Sub-group discovery, contrast set mining, formal concept analysis, exploration hypercubique...

Un domaine de recherches très actif et fécond

Vague du « data science » :

- Développement d'outils mathématiques et algorithmiques très pointus pour analyser les données biologiques
- Floraison de nouvelles méthodes et algorithmes disponibles pour les chercheurs sur le Web (packages R, Matlab...)
- Des équipes françaises en pointe (ENS, INRIA, Mines...)

Quelques développements récents :

- Sparse learning : régularisation et réduction de dimension pour données de haute dimension (Elastic net, SPLS...)
- Topological data analysis (Ayasdi)
- Co-expression networks (Aracne)
- Deep learning (nouvelle génération de réseaux de neurones)
- Tensor decomposition
- ...

Big data dans la santé, quel progrès médical en attendre ?

41

Colloque XMP Biotech, 17 mars 2014

Les difficultés à surmonter restent considérables

Corrélation n'est pas causalité...

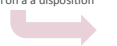
OPEN ACCESS. Preprint available online.

PLOS COMPUTATIONAL BIOLOGY

Le data mining permet de « dénichier » des explications aux phénomènes qui sont inaccessibles aux approches statistiques classiques

Mais :

- Forte dépendance au contexte et à la qualité des données analysées
- Bases de données patients petites (nb de cas) vs rapport au nombre de variables : risque élevé de « **faux positifs** » (sur-apprentissage, multiple testing)
- Certaines lois de causalité (interactions) échappent
- On ne trouve ce qu'il y a dans les données que l'on a à disposition



Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet¹, Jacques E. Dumont², Vincent Detours^{3,4,5}¹ MEDA-GISE, Université Libre de Bruxelles (ULB), Brussels, Belgium, ² BRBIB, Université Libre de Bruxelles (ULB), Campus Doctus, Brussels, Belgium, ³ BRBIB, Université Libre de Bruxelles (ULB), Campus Doctus, Brussels, Belgium

Abstract

Bridging the gap between animal or in vitro models and human disease is essential in medical research. Researchers often neglect that a biological mechanism is relevant to human cancer from the statistical association of a gene expression marker to a signature of this mechanism, that was discovered in an experimental system, with disease outcome in humans. We examined this argument for breast cancer. Surprisingly, we found that gene expression signatures—associated to cancer—of the effect of postprandial laughter, of mice social defeat and of skin fibroblast localization were all significantly associated with breast cancer outcome. We next compared 47 published breast cancer outcome signatures to signatures made of random genes. Twenty-eight of them (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worse predictors than the median random signature. More than 50% of random signatures (100 genes) were significant outcome predictors. We next defined a metagene, called meta-PCHA, by selecting the 1% genes most positively correlated with published marker PCHA in a comparison of normal versus metastatic. Adjusting breast cancer expression data for meta-PCHA altogether almost entirely the outcome association of published and random signatures. We also found that, in the absence of adjustment, the hazard ratio of outcome association of a signature strongly correlated with meta-PCHA, $HR \sim 0.8$. This relation also applied to single-gene expression markers. Moreover, ~20% of the breast cancer transcriptome was correlated with meta-PCHA. A caveat was that upregulating cell cycle genes out of a signature failed to rule out the confounding effect of proliferation. Hence, it is questionable to suggest that a mechanism is relevant to human breast cancer from the finding that a gene expression marker for this mechanism predicts human breast cancer outcome, because most markers do. The methods we present help to overcome this problem.

Axes de progrès :

- Augmentation du nombre d'observations
- Traçabilité et qualité des données source
- Progrès dans les technologies d'analyse (recherche mathématique)
- Réplication et validation fonctionnelle systématique des résultats obtenus

Big data dans la santé, quel progrès médical en attendre ?

42

Colloque XMP Biotech, 17 mars 2014

Conclusion

En résumé

Une révolution économique, technologique, scientifique est à l'œuvre dans le domaine de la médecine et de la santé.

Cette révolution est portée par la mise à disposition d'un volume et d'une variété de données sans commune mesure avec ce que nous avons connu dans le passé proche.

Etre capable de tirer de la valeur de ces données est un enjeu majeur.

L'analyse à valeur ajoutée des données biologiques et médicales reste un challenge :

- Multi-causalité
- Populations hétérogènes
- Données de qualité hétérogène et parfois faible
- Obstacles à la consolidation des données provenant de multiples sources

Les technologies d'intelligence artificielle et de data mining joueront un rôle clé dans les progrès à venir.

Les progrès dans la gestion du processus d'intégration et de gestion des données et l'augmentation du nombre de patients dans les bases permettront à ces technologies de donner leur plein potentiel.

Big data dans la santé, quel progrès médical en attendre ?

44

Colloque XMP Biotech, 17 mars 2014

Quels défis à relever pour un ingénieur passionné par la Biotech et la santé ?

Technologiques :

- nouvelles technologies de next generation sequencing (microfluidique)
- Nouvelles techniques d'imagerie
- Mobile health devices

Traitement du signal, traitement des données

- Des "raw data" à l'information

Informatiques

- Cloud computing, stockage, consolidation, intégration, fédération, analyse de volumes énormes de données éclatées

Data science, analyse de données (Détection des signaux faibles, Prédiction)

- Recherche fondamentale (améliorer les méthodes et les algorithmes)
- Recherche appliquée (discovery avec des objectifs biomédicaux)

Ingénierie de projets complexes

- Consolidation, intégration, analyse de données multi-source...

Multidisciplinarité, esprit entrepreneurial, créativité, sciences de l'ingénieur, compétences mathématiques et scientifiques...

Big data dans la santé, quel progrès médical en attendre ?

45

Colloque XMP Biotech, 17 mars 2014