HEALTHCARE
**DATA** INSTITUTE

**INTERNATIONAL THINK TANK**
DEDICATED TO BIG DATA IN HEALTHCARE

# BIG DATA AND PREVENTION
## FROM PREDICTION TO DEMONSTRATION

NOVEMBER 2016

# BIG DATA AND PREVENTION **FROM PREDICTION TO** DEMONSTRATION

Acknowledgements to the authors of this White Paper:

- **Isabelle Barbier-Feraud**, Pagamon
- **Jeanne Bossi Malafosse**, Lawyer
- **Patrice Bouexel**, TeraData
- **Catherine Commaille-Chapus**, Open Health Company
- **Anne Gimalac**, Aston Life Sciences
- **Guillaume Jeannerod**, Epiconcept
- **Magali Léo**, CISS
- **Bruno Leroy**, Sanofi
- **Bernard Nordlinger**, AP-HP, Académie de Médecine
- **Michel Paoli**, InterMutuelles Assistance
- **Pablo Prados**, Sanofi
- **Jean-Yves Robin**, Open Health Company

# INTRODUCTION

Taking action to prevent epidemics, identifying risk factors as soon as possible in order to avoid diseases or slow their development, predicting predisposition to particular infections from birth, and so on. These are among the goals stated time and time again by scientists and representatives of the medical community.

Up to now, the response to these issues has involved the use of traditional research methods, which have been and are still largely based on the notion of reproducibility: using a hypothesis that serves as the basis for the collection of data, an observation makes it possible to develop a relationship among those data and to deduce new hypotheses or new treatments, which will then be reproduced in similar cases.

For a number of years, this research has been enhanced by other very diverse sources that also take into consideration the environment in which the individuals are living: environmental data, socio-professional data, etc.

This situation is now being re-examined by the sudden emergence of new data management techniques, which must adapt to the massive, exponential production of data characterised by the phenomenon of Big Data (we will come back to a definition of this term a bit later).

Can this massive production of data have an effect on disease prevention? Does Big Data have a role to play, among other technologies, in improving the health of the population and thus providing

politicians with the right information at the right time so as to enable the definition of a more intelligent public health policy?

The development of digital technologies over recent years has also resulted in a veritable 'datification' of our society. Data are everywhere, and constitute a valuable raw material in the development of new knowledge and a major global factor for growth for a number of countries.

This development is also characterised in the health sector by the increasing computerisation of the professional sector, in particular activities associated with care and prevention, life sciences research, and health systems management, and by the growing involvement of patients.

The phenomena of the mobility and development of connected medical objects and devices are also contributing to the exponential growth of the volume of data produced, and the phenomenon of Big Data merely reflects the specificities involved in the processing of large volumes of data, with the associated requirements for speed of processing, uniformity of the data and specific value creation.

Health data therefore represent a unique tool and a potential to create value, with the realisation of that potential depending on the ability of the various nations to organise the development of an ecosystem that facilitates the use of the data while guaranteeing compliance with privacy requirements and the confidentiality of personal data.

Indeed, governments are currently being asked to address major issues in which the processing of health data can play and is already playing an essential role: questions of public health, questions of quality of care, questions of transparency and health democracy, questions of efficiencies within health systems in a general context of increasing health expenditure, and questions of innovation and growth in areas as varied and significant as personal medicine or precision medicine and information technologies.

More than two years ago, the Healthcare Data Institute initiated a process of reflection and discussion on the subject of Big Data and health, and has put together a working group specially dedicated

to the subject of Big Data and prevention. Its goal was to provide a few examples illustrating the role that Big Data can play in health prevention, to identify the factors that promote its development and those that frustrate it, and to propose certain areas for investigation that could be valuable for decision-making, both private and public.

The work described in this White Paper is the result of the discussions and activities undertaken by individuals from a range of different private and public professional contexts. It is clearly not intended to be exhaustive on this subject but, rather, to draw attention to domains that are currently integrating new areas of potential for Big Data and, at the same time, to highlight certain problems of varying types. ■

# DEFINITIONS

In order to present this work, there was a need to agree on two important definitions - the definition of Big Data and the definition of prevention.

**Big Data** describes data sets that become so large that they are difficult to process using only database management tools or traditional information management tools. Big Data also describes all of the technologies, infrastructures and services that make it possible to collect, store and analyse data gathered and produced in increasing quantities, using automated processing operations and artificial intelligence technologies[1].

It is standard practice to illustrate this definition using the image of the '3Vs' that characterise Big Data: the explosion of the **volume** of the data, the **variety** of the structured and unstructured data produced by multiple sources, and **the velocity** of the information and simultaneous processing speed.

A fourth V should be added to these three: the **value** represented by these data for the entity or individual. Data have become the raw material of the digital universe.

It is easy to see the confusion that often arises between the term Big Data and the usage of data analysis methods such as Data Mining. As a result of this confusion, the term Big Data is used very extensively to describe all new data analysis methods.

The working group however deliberately made the decision to use this broad definition of the term Big Data in order to adapt the scope of its analysis to the actual realities of the subject: the question we need to address today is more about knowing how to create intelligent databases in order to manage Big Data.

So, in this White Paper, the term Big Data will be used flexibly to group together all of the techniques and methods currently used to

1. Report from the Institut Montaigne on Big Data and Connected Objects, 2015.

analyse the increasingly large volumes of data produced by a variety of sources.

With regard to **prevention**, which is also a very vast term, we have used the definition provided by the World Health Organisation (WHO). According to that institution, prevention is '*all of the measures intended to avoid or reduce the number or severity of diseases or accidents*'.

Three types of prevention are then identified:

• Primary prevention, which covers *'all of the actions intended to reduce the incidence of a disease, and therefore reduce the frequency of new cases'*. This involves the use of individual prevention measures (personal hygiene, food, etc.) and/or collective prevention measures (distribution of potable water, vaccination, etc.).

• Secondary prevention, which covers *'all actions intended to reduce the prevalence of a disease and therefore reduce its duration'*. This includes screening and the treatment of new outbreaks.

• Tertiary prevention, which concerns '*all actions intended to reduce the prevalence of chronic disabilities or relapses in the population and therefore reduce the functional disabilities caused by the disease*'. The aim of these strategies is to promote social and professional reintegration following the disease. This definition extends the notion of prevention to cover rehabilitation treatment. ■

# TABLE OF CONTENTS

The subjects addressed in this White Paper by means of a series of articles are organised around several horizontal and transverse themes. Each article can be read independently from the others.

**I.** An initial theme addresses the role of Big Data and personalised prevention on the basis of the issue of genomics.

**II.** The second theme discusses universal population prevention and illustrates the possibilities for the management of real-life data to benefit public health and the prevention of health risks.

**III.** The third theme selected covers the essential subject of the development of the technological landscape as a key success factor for 'making the data speak'.

**IV.** The fourth theme addresses the question of the definition of a new paradigm for the stakeholders involved.

The choice of these four subjects has made it possible for the members of the working group to identify the common, cross-cutting problems in relation to each theme that appeared to warrant analysis and could serve as the basis for recommendations.

This is covered in Article **V**, in relation to the important issue of the anonymous or pseudonymised nature of the data used and the appropriate degree of aggregation to be applied, and Article **VI** on ethical and legal aspects, which are covered in this document essentially through the theme of patients and making data accessible.

# ARTICLE 1
# BIG DATA AND GENOMICS: WHAT FUTURE DO THEY HAVE IN THE TREATMENT OF CANCER?

## 1. CONTEXT & TECHNOLOGY

It has been said that we are entering the era of '6P Medicine' (Personalised, Precision, Participative, Preventive, Predictive and Patient-oriented) based on a genomic and molecular approach to disease that provides an opportunity to:

- better understand pathological mechanisms;
- identify new therapeutic targets;
- identify risk factors;
- support diagnosis and medical decision-making;
- personalise treatment.

What is this?

For a given condition, the sequencing of all or part of a patient's genome makes it possible to identify the orthographic differences (referred to as variants) compared to the genome of individuals who are not suffering from the condition in question. Over recent years, considerable technological advances in the area of high-throughput sequencing (genome decoding method) have made it possible to reduce the associated times and costs and to facilitate the storage and analysis of the mass genomic data. This could suggest that the sequencing of the entire genome (and not just certain target sections) for individuals and patients could soon be a matter of routine, paving the way for applications in the domains of both research and treatment.

But where actually are we at the moment? Can we really apply Big Data to the realm of genomics? Will we reach that point soon or will it take us longer? What will be the outcome and which obstacles will we need to overcome to get there?

In the remainder of this section, we have chosen to illustrate the example of cancer.

## **2.** SCIENTIFIC RESEARCH AND THE TREATMENT OF CANCER TODAY

Considerable progress is currently being achieved in the understanding and treatment of cancer, based both on a better understanding of the impact of environmental factors, on clinical observation, on clinical trials of new pharmaceutical treatments, and on progress in immunotherapy (in particular for melanoma) and biopsies on fluids (and not merely tissues).

We are also seeing the development of the genome analysis of tumours themselves and of patient genomes. By way of example, at the French National Cancer Institute (INCA), 70,000 new patients each year are having their genome subjected to the targeted sequencing of certain regions, followed by routine analysis in order to identify any molecular anomalies and to personalise their treatment in order to increase efficacy and reduce side effects. It is expected that whole-genome sequencing will be routine by 2020, and this should therefore facilitate a systemic understanding of the disease.

Furthermore, we are seeing a change in the way new molecules are being developed from academic and industrial research, and in how the efficacy of those molecules is demonstrated during regulated clinical trials. Patients recruited for Phases I, II and III clinical trials can now be selected on the basis of their genome profile. Previous Phase III studies are, in particular, being reanalysed in the light of the discovery of new **biomarkers** (certain specific genetic markers expressed by the tumour) in order to evaluate the benefit of targeted therapies for tumours in patients with a genome profile associated with a better response to the treatment, in other words those who make it possible to predict how a tumour will respond to a given treatment. The idea is to have access to data that enable the administration to a patient of the treatment that is best suited to his or her disease, in order to increase their chance of recovery. It has now, in fact, been clearly established that not all patients suffering from

cancer react to a treatment in the same way. Response depends on both the specific characteristics of the patient and the specific characteristics of the tumour. We are therefore researching biomarkers correlated to a better response to a treatment and biomarkers that make it possible to estimate the risk of relapse.

**Prognostic biomarkers** are biomarkers that allow us to better predict the lifetime or lifetime without relapse for the patient, as another means of improving treatment.

In the case of colorectal cancer, we have been able to establish that only patients whose tumours have a normal (non-mutated) version of a gene called RAS are able to benefit from the effect of cetuximab and panitumumab, two new medicinal products (monoclonal antibodies) for targeted therapy. In this case, **predictive biomarkers** are biomarkers that make it possible for us to better predict the efficacy of a treatment.

But here, we still cannot really consider this to be a Big Data approach to genomics, in particular in terms of mass data or diversity of sources. Although we now know how to cross-reference the genetic information characterising the tumour and the response to a therapy for several hundreds or thousands of individuals, the selection of treatment based on genome sequencing has still not become systematically routine in medical practice.

At present, the application of Big Data in genomics only occupies a very small space among the methods for research and treatment of disease, and cancer in particular.

There are however certain isolated initiatives that are moving in this direction, including the following examples (non-exhaustive list): in England (*The 100,000 genomes project*); in Germany (systematic genotyping of tumours at the *DKFZ* - German Cancer Research Centre in Heidelberg, a model to be applied); in the United States (*CancerLinQ*, *Cancer Moonshot*); in France (*France Médecine Génomique 2025*, INSERM/Quest Diagnostics '*BRCA Share*' to improve diagnosis of predisposition to breast and ovarian cancer); in Switzerland (*Personalised Health*); in Europe (*SPECTAcolor*, *Million European Genomes Alliance*, *Sophia Genetics**), etc.

*It is interesting to observe a new trend for the application of Big Data, at the boundary between biotechnology, genomics, digital health and artificial intelligence. Using sophisticated algorithms, 170 hospitals in 28 European countries are currently using the services associated with analysis and interpretation of the genome profiles of their cancer patients (on the basis of pooled data) in order to support decision-making by their doctors in relation to the degree of pathogenicity of their genetic mutations. The goal is not for the algorithm to replace diagnosis by a physician but, rather, to accelerate the diagnosis of 'simple' cases in order to free up the doctor's time for the diagnosis of more complex cases that involve the accumulation of a range of different genetic mutations.

## **3.** HOW CAN MASS GENOMIC DATA ANALYSIS CONTRIBUTE IN THIS MECHANISM? WHAT CAN WE EXPECT FROM BIG DATA?

In epidemiology, the analysis of mass data, cross-referencing the genetic/genomic data about patients and their tumours with external environmental data (geography, diet, atmospheric pollution, UV radiation, etc.), connected health data, etc., would considerably improve the effectiveness of current treatments, making it possible to better identify the risk factors that influence cancer survival and, therefore, contribute to their prevention.

Any progress in this area will depend on the possibility of cross-referencing clinical and genetic data on tumours and patients, and therefore on the creation of biobanks and prospective and retrospective databases, and access to those tools that is unrestricted but strictly regulated in terms of protection of medical data. **We will have gen-**

uinely entered the era of Big Data in genomics when we are able to accumulate genomic and diagnostic data on large numbers of patients, clinical data before/during/after treatment, etc., that make it possible to define potential correlations between genotypes and phenotypes and thus to predict the influence of environmental factors and susceptibility to diseases or therapeutic incidents. These

Prof. Pierre-Laurent Puig, speaking before the Working Group, physician in the Biochemistry Department of Hôpital Européen Georges Pompidou (Paris) and Director of the INSERM UMR-S775 Research Unit at the Université de Médecine Paris Descartes. Research subject: Colorectal cancers.

*'Today, the generation of genomic data is easy and not costly. In terms of application, we can no longer dissociate scientific research from medical treatment, because we are analysing the associated raw sequencing data in order to characterise a specific medical condition and adapt the treatment of patients suffering from that condition. However, in France, we are facing a regulatory obstacle (not technological or financial) associated with the protection of health data, which does not authorise the cross-referencing and correlation of genomic data with clinical data from medical files, and does not authorise the use of data for purposes other than those for which they have been generated and collected. The databases are created retrospectively. The prospective approach, bringing together data collected without a central purpose, will be possible when we are able to put in place electronic medical records that are centralised from the outset and created for the purposes of management and research without any pre-established purpose.'*

prospective and retrospective databases will be dedicated to research, diagnosis and treatment, and could make it possible to address the questions raised, eliminating the requirement associated with the initial purpose for which the data are collected.

## **4.** WHAT ARE THE OBSTACLES TO OVERCOME IN ACHIEVING 6P MEDICINE BASED IN PARTICULAR ON BIG DATA AND MOLECULAR DIAGNOSIS?

Please find listed below what we believe to be the primary technological, organisational, political and economic challenges.

### **A.** TECHNOLOGICAL CHALLENGES

The software and algorithm market requires:

1) opening up to digitised/digital data (essential for traceability, even if this creates new risks);
2) professional knowledge;
3) standardisation, repeatability.

It is therefore necessary to:

• Develop standardised electronic patient files incorporating genomic data and traditional clinical data.

• Create the ecosystem that makes it possible to bring together the skills of engineers, IT technicians, molecular geneticists, geneticists, etc.

• Develop algorithms, applications and processes to ensure confidentiality and security of data so as to enable medical practitioners to provide a clear opinion, diagnosis and treatment regimen.

• Have the algorithms validated by all of these experts so as to guarantee that they are robust, easy to use and applicable (IBM Watson is only applicable for American medical practice!).

But also to:

• Put in place a multidisciplinary network to address the exponential

requirements for high-throughput sequencing, collection, storage, large-scale analysis, and decoding of the databases to ensure that they can be understood by the practitioner.

## **B.** ORGANISATIONAL CHALLENGES

• Guarantee the confidentiality of the genetic data and ensure compliance with a regulatory framework so as to avoid discrimination.

• Prepare doctors for the coming genome revolution, train them in changing practices, prepare them for the decompartmentalisation of medical specialisms (as the disease will no longer be viewed in terms of its location within the body but, rather, at a molecular and cellular level).

• Develop new tools with the practitioners themselves in order to make them ambassadors contributing to the development of medical practice and the design of the medicine of the future.

• Implement work in synergy with other professions (bio-IT, statistics, etc.) to determine the most appropriate protocol.

• Develop the doctor/patient relationship to support patients excluded from sub-groups, for which a medicinal product has been targeted, to support patients with significant genetic predisposition, to educate patients in new preventive behaviours, and to ensure a better understanding by the patient, thus enabling informed consent.

• Reorganise scientific and technological research on the basis of the development of medical and clinical practice, consider new public/private arrangements, etc.

## **C.** POLITICAL AND ECONOMIC CHALLENGES

• Develop the health system, determine the conditions and levels of treatment for health insurance.

• Create an industry sector around genome and digital health, to mobilise the sequencing, data storage and scientific instrument industry, and to design a new model for the pharmaceutical industry.

• Bring together stakeholders such as the pharmaceutical industry, with teaching hospital institutions, as early as possible in the value chain, in order to provide access to genomes from cohorts of patients and identify the biomarkers associated with various conditions.

# ARTICLE II
# DATA AND POPULATION HEALTH PREVENTION

In health as in other fields, technological progress has caused an explosion in the quantity of information collected every moment. The use and analysis of these growing volumes of data available from a variety of different sources that collect them for various reasons are a source of considerable hope and progress in terms of scientific advances, and create new opportunities for population health prevention and the management of health risks.

Population health prevention is one of the roles of public health, defined as the 'science and art of promoting health, preventing disease, and prolonging life through the organised efforts of society[2]'. In this interpretation, prevention covers all measures intended to avoid the occurrence of diseases and accidents or to reduce their number, severity and consequences.

Big Data are based on the ability to collect, aggregate and process data originating from a wide range of heterogeneous sources. The variety of data collected (structured, unstructured, etc.) and sources (public/private databases, medical and administrative data, health data/environmental data, etc.), their volume, and the velocity with which they are gathered and processed are at the heart of approaches based on Big Data. The spread of the use of connected objects and the development of geolocation practices are contributing to the significant growth in the volume of available data. The analysis of these large-scale data, which is made possible both by the increasing digitisation of activities within society and by the much-enhanced capacities for storage and processing, is revolutionising traditional approaches to prevention.

The analysis of data has always been at the heart of our understanding of health phenomena. The knowledge gained by the public

2. D Nutbeam, WHO, 1998.

authorities, which is necessary for action to be taken, is historically based on activities involving epidemiological observation, which are intended to observe and measure the health-related events affecting a given population, to explain them, and to determine the impact of the measures taken to address them.

What is new today using the techniques associated with Big Data is this new ability to aggregate and process massive volumes of data originating from different sources, coupled with the fact that there is now a considerable bank of digital data available, both in the health sphere (treatment data, medical and administrative data, data produced by patients themselves using connected objects, etc.) and more widely about socioeconomic, cultural and environmental conditions, which are the key determinants for population health, such that there is no longer necessarily any need to produce these data specifically. The European Regulation on the protection of personal data, adopted on 27 April 2016, thus enshrines, for the first time, the concept of 'compatible purpose', authorising the reuse of data for a purpose other than the purpose for which they were initially collected, provided that the intended purpose is compatible with the initial purpose.

While up to this point, health-related studies have required the collection of ad hoc data, on the basis of criteria used to populate databases, often over very long periods, data produced during treatment activities or for the purpose of reimbursement of costs, or even by patients themselves using connected objects or social networks, are now available and constitute an almost inexhaustible source for the identification of disease risk factors, health security or epidemiology.

The processing and analysis of these still largely under-used big data provide limitless opportunities for knowledge generation. These analyses can contribute to the prevention of diseases and epidemics and to surveillance and medical supervision, making it possible to better understand the socioeconomic and environmental determinants for population health, to detect unusual health events that could represent public health alerts, and where necessary to establish links with exposure-related factors. They can make it possible to target prevention efforts towards the population groups for which these measures are most effective, and can contribute to the

monitoring and evaluation of public health actions.

The following sections illustrate the variety of possibilities provided by Big Data in terms of population health prevention, supervision and health security.

## 1. OPPORTUNITIES FOR THE USE OF BIG DATA IN PHARMACOVIGILANCE

In a country that is classified among the biggest consumers of medicinal products in Europe, the question of the adverse effects associated with their consumption and the monitoring of correct use is increasingly becoming a major public health issue. The recent health crises (Mediator, 3rd and 4th generation pills, Depakine, etc.), the impact of which is clearly considerable in terms of public health and in financial terms, illustrate the difficulties that the health authorities and the national community must address in relation to the misuse of drugs that is more extensive than in the past, combined with a lack of precise information about the conditions for use of drugs[3]. Allowing misuse to continue or discovering the serious effects of this phenomenon ten years after the fact is both dramatic in terms of population health and extremely costly for the community. In European terms, the cost of adverse effects linked to drugs is estimated as 197,000 deaths per year in the EU. The financial impact is valued at 79 billion euros per year[4].

The monitoring of the safety of drugs is based essentially on the evaluation actions that can be undertaken following market introduction and dissemination within the general population, in population groups that are broader and more diverse in terms of their characteristics than those included in clinical trials, and for which the use of the drug (treatment period, posology, observance, etc.) is itself

3. B. Bégaud, D. Costagliola, *Report on the monitoring and promotion of the correct use of drugs in France [in French],* pharmaco-monitoring project entrusted by the Minister of Social Affairs and Health, Ms Marisol Touraine, 26 February 2013.

4. H. Pontes, M. Clément, V. Rollason, *Safety Signal Detection: The Relevance of Literature Review*, Spinger International Publishing Switzerland 2014.

much more variable[5]. The analysis of the data from patient cohorts or medical and financial databases over the long term can make it possible to detect signs and establish links between the occurrence of a health event and exposure to a given treatment, and can trigger an alert as to the possible adverse effects of drugs or off-label use not in accordance with the indication for which a given health product has been awarded a marketing authorisation.

The speed of access to the relevant knowledge is absolutely key. In the case that serious adverse events are detected, the periods in years that can elapse between the suspicion of a phenomenon and the establishment of a sufficient level of proof can be highly damaging. Big Data-type search techniques can be applied to support these analyses and can enable the early detection of warning signs that are decisive in public health terms.

## **2.** BIG DATA IN THE PREVENTION AND MONITORING OF EPIDEMICS

The availability of large-scale data from different sources, in real-time or near-real-time, means that information can be obtained about the health status of a population in a given geographical zone. The cross-referencing and analysis of these data, combined with mathematical modelling techniques, can make it possible to identify trend shifts that herald an increase in the incidence of diseases or behaviours and suggest that a probable change in the health status of that population can be expected.

Thus, alongside traditional information gathering and alert mechanisms for infectious diseases and epidemics (such as the Sentinelles network in France, made up of doctors throughout the country who report the cases observed each week for a certain number of transmissible diseases), epidemiological models are being developed to monitor the space-time propagation of health-related phenomena such as epidemics through the use of large-scale data.

5. J.L. Faillie, F. Montastruc, J.L. Montastruc, A. Pariente, *The contribution of pharmacoepidemiology to pharmacovigilance*, 2016.

A little over ten or so years ago, the use of Big Data was initiated within the new public health agency to bring together the InVS, the INPES and the EPRUS, which was entrusted with the task of protecting the health of the French population. The associated processes were based above all on the use of data from defined, known data sources.

Of the 3Vs frequently associated with Big Data, Variety is most important.

By using stable, validated data sources such as PMSI, SNIIRAM, CepiDC or data from cancer registers, scientifically validated indicators that have been shown to be robust over time have been developed.

SurSaUD (the system for the health surveillance of emergency departments and deaths), the monitoring system put in place following the 2003 heatwave and based on data from A&E departments, SOS Médecins, mortality data and death certificates, is just one example.

These indicators make it possible to produce the BQA (Daily Alert Bulletin) intended for the health authorities, including the Minister of Health.

A system of this kind is highly valuable in detecting unexpected health-related events, estimating the impact of an environmental or societal event, monitoring medical conditions where no events have been detected, or enabling the early detection of a predefined health-related event, such as a seasonal epidemic, and in measuring the impact and consequences of that event.

While the various systems put in place are effective in detecting an event or confirming a trend, they have limited or nil capacity to explain the associated causality within the deadlines imposed by the need to respond to these situations and by political pressure.

For information purposes, in the Île-de-France Region in March 2014, an increase in asthma was observed and attributed initially to the peak in pollution that was rife at that time. Subsequently, it became clear that a larger quantity of pollen than usual was circulating in the air at that time and that the pollution was merely a third factor.

The example of Google Flu Trend, which was intended to monitor

the development of the flu around the world on the basis of an analysis of certain key words within the search engine, has been repeatedly highlighted. A comparative analysis of detection methods with other methods for monitoring the propagation of the flu syndrome has shown the limitations of this model in respect of determining the extent of the phenomenon. Google was forced to reintegrate data originating from the American Centres for Disease Control (CDC) and the service has since been closed.

The use of data from mobile telephones is a very interesting example, as these data can in fact make it possible to provide a very precise description of the effects that can be generated by public gatherings and movements on the propagation of an epidemic such as cholera or malaria. An understanding of these phenomena, which enable the identification of 'transmission points' for an epidemic, has huge potential in terms of contributing to the eradication of infectious diseases[6].

Other initiatives, such as an analysis of the browsing frequency for certain Wikipedia pages (such as the page on bronchiolitis) or an increase in keywords on Twitter, enable the detection of weak signs.

While it is important, as these examples show, to remain cautious in terms of the validation of these analyses, which require observations gathered on the basis of long case series over time and over distance, and while these models have not yet shown their full potential, the fact remains that they offer excellent prospects in terms of the prevention and control of pathologies, and especially infectious conditions.

The power of the algorithms and an increase in sources of data over time will provide a guarantee that every health-related event will be detectable.

However, it is and will always be essential to allow for the time and the medical expertise provided by the men and women in the field to triage the vast quantity of events detected and understand and

6. Flavio Finger, Tina Genolet, Lorenzo Mari, Guillaume Constantin de Magny, Noël Magloire Manga, Andrea Rinaldo and Enrico Bertuzzoa, *Mobile phone Data highlights the role of mass gatherings in the spreading of cholera outbreaks.*

ensure a reliable matching between a health-related event and its cause. This is not necessarily compatible with the information society that is developing at the same speed as these technologies.

## **3.** CROSS-REFERENCING OF CLINICAL AND SOCIO-DEMOGRAPHIC DATA: MOVING TOWARDS AN IMPROVED PREDICTIVE FACTOR FOR TREATMENT COMPLIANCE?

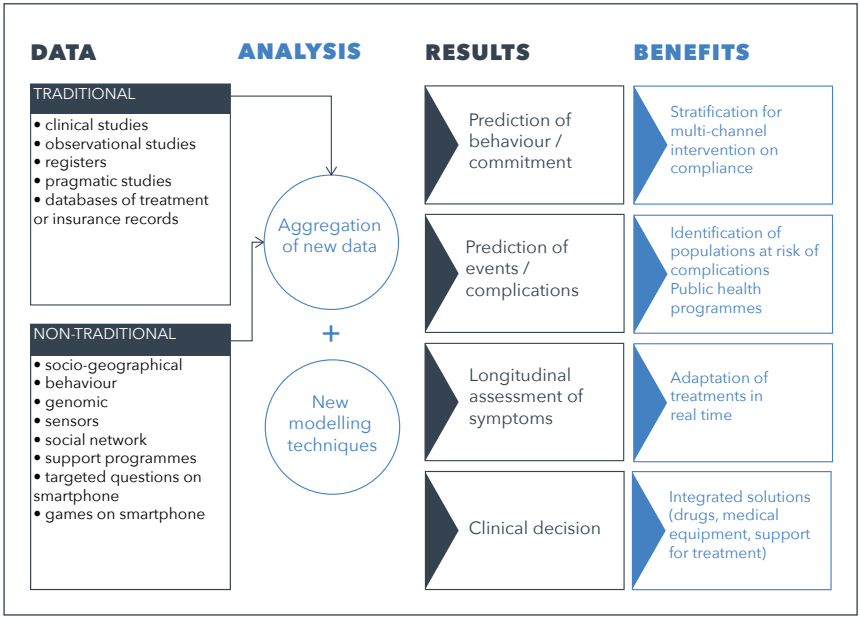
Socio-geographical data, traditionally not used in the health sphere, could be integrated into the design of more personalised public health intervention programmes, by means of modelling using advanced machine-learning techniques.

These techniques make it possible to develop models for predicting treatment compliance and the onset of complications such as acute cardiovascular events or episodes of severe hypoglycaemia in patients suffering from chronic diseases such as diabetes. The public health intervention programmes could then be optimised by giving priority to the use of certain resources for the population categories most exposed to these complications, based on their risk of non-compliance with treatment assessed on the basis of certain geographical characteristics or characteristics of socioeconomic behaviour. Models similar to those used in mass consumption to predict purchasing behaviour are currently being tested with this in mind. Most systems of recommendations make suggestions of personalised choices based on the experience drawn from previous purchasing behaviours. These approaches have not yet been used extensively in public health and represent a possible source of improvements in the quality of care.

This approach firstly requires that reimbursement data be linked anonymously to demographic medical data and data on the course of the disease, in order to identify the predictive factors for treatment compliance and for complications associated with the diseases concerned. The socioeconomic and geographical data are then added to the analysis to compare or improve that prediction and thus guide

the intensification or simplification of the intervention programmes.

Numerous socioeconomic and geographical data are now available freely in certain countries such as the United States (census data, consumer expenditure figures, national salary study, food environment, housing survey, etc.). One critical point, nonetheless, is the size of the region concerned, given the need to ensure the anonymity of the medical data. Data on patient behaviour can also be obtained using digital applications, enabling, for example, the monitoring of social and physical activity using a smartphone, collected directly or using simplified questionnaires administered regularly.




The types of data that can be analysed, the associated opportunities and the possible impact on patient management are summarised in the graphic above.

# 4. BIG DATA AND THE PERSONALISATION OF PREVENTION

The field of prevention covers both collective actions intended to protect the health of persons through the development of a physical and social environment that is favourable for health (health monitoring, vaccination, screening for certain diseases, preservation of air and water quality, etc.) and the promotion of individual behaviours that are favourable for health (prevention of alcoholism and smoking, promotion of physical activity and nutritional advice ('eat well, be active'), etc.).

Prevention activities are therefore aimed at the entire population and exist in all places in the lives of individuals: at home, at work, at school, in the city, etc.

Historically, prevention programmes have been implemented by means of large-scale actions most often referred to as 'campaigns':

- vaccination campaign;
- screening campaign;
- campaign to raise awareness about at-risk behaviours, etc.

These campaigns, conducted by the organisations responsible for population health, are very non-specific, in the sense that they are intended to reach the largest possible number of people, often on the basis of cost/effectiveness analyses. The individual benefit is then a consequence of the benefit for the population:

- Providing information for the entire population by means of a radio or televised message in order to launch a vaccination campaign against the flu while focusing on the elderly population.

- Proposing a mammogram every two years in all cases for all women aged 50 to 74 years.

In the two examples cited above, we can see that the campaigns are aimed at a very broad audience in order to increase the impact on the population. There is little targeting, and the persons concerned are supposed to recognise themselves in the mechanism put in place.

For screening campaigns, in the case of a positive test, medical

follow-up and potential treatment will be customised as part of a personalised care path downstream from the screening activities.

Today, the availability of the data and the intelligence that can be obtained by cross-referencing those data make it possible to aim for the largest possible number of individuals while targeting the messages, redirecting people who are at risk or excluding those not affected.

By using the same techniques as those used on the internet for advertising, it is possible to get the right message to the right person, taking into account various items of 'highly personal' information: medical history, drug consumption, genetic information or behaviour.

We can cite the change that is currently taking place in the breast cancer screening programme. Initially, this was a vertical programme dedicated to an identified population: women aged 50 to 74 years who were offered a screening test every two years. This programme will be optimised as a result of the availability of genetic data from women with hereditary predispositions to breast cancer.

The mechanism will be adapted to create a system where systematic, mass screening becomes 'personalised' mass screening that takes into account the risk factors specific to each woman, with appropriate personalised monitoring plans. The French National Cancer Institute (INCA) is currently working on the introduction of these programmes.

The more these data are available, the more specifically the campaign will be targeted and the more precise and tailored the 'prevention proposal' will be.

These possibilities offered by Big Data should ensure a significant improvement in public health programmes.

This promise of greater efficiency and substantial savings has been clearly understood and integrated by bodies such as suppliers of connected objects in the area of well-being, the members of GAFAM (Google, Apple, Facebook, Amazon and Microsoft), the mutual insurance organisations and, more recently, the health insurance sector (the 'active health' programme).

Individual coaching can be provided by artificial intelligence systems on the basis of behaviour. In the context of the mutual insurance organisations, 'good behaviour' can be rewarded by presentation of a gift or even a change in the amount of the premium in certain countries such as the United States.

For the pooling of risks, the founding principle underlying public health programmes and health insurance systems in the broad sense, there will be an increase in the temptation to use models that will identify 'bad behaviours' by individuals or propose that at-risk people be excluded from the system.

It is easy to imagine that we could see a move towards elitist, behaviourist models promoted by entities managing data for the purposes of creating profit for society and for themselves.

The question of the roles of the various parties is being asked more than ever before: governments, institutions responsible for population health, experts, companies and bodies that own the data.

For these prevention programmes, the role and responsibilities of each must be clearly stated and applied so as to avoid any abuses, which would ultimately conflict with their reason for existing.

## **5.** CONCLUSION: METHODOLOGICAL AND ETHICAL FACTORS

The cross-referencing of an increasing quantity of health-related and other data results in the creation of hypotheses relating to health and the associated environmental and socioeconomic determinants, and could make it possible to provide access to new knowledge. These data originate from a wide range of structured and unstructured databases, using measurement methods that vary depending on the database and the year, missing data, sources, subject origins and methods of collection that are extremely varied. The question of method in this context and in the context of controlling information is becoming absolutely central for our public system health[7].

7. C. Dimeglio, C. Delpierre, N. Savy, Y. Lang, *Big Data and public health: more than ever,*

Similarly, the availability of the data and the possibilities that this entails in terms of assessing the effectiveness of public health actions should lead to new responsibilities for the stakeholders in the system. The monitoring of data in real time can make it possible to assess and measure the effectiveness of actions taken and to adjust them where appropriate. However, it is becoming impossible to ignore the absence of results. This is the case, for example, with the failure of the campaigns in relation to vaccination against seasonal flu that are repeated in exactly the same way year after year, while the results in terms of population coverage remain very much below the recommendations issued by the WHO. This has consequences in terms of increased mortality associated with the epidemic each year.

The data, where they are available and able to provide clarification about the efficiency of the action, detecting signs and monitoring epidemics and health crises, can no longer remain inert. Their access must be open to multiple stakeholders for public health purposes, fostering permeability between the entities responsible for health policy and the world of data sciences, developing modern modes of regulation and cooperation between the political sphere and the stakeholders in the economic world, the drivers of innovation, and supporting the accessibility and use of health data. This is to take place under conditions that guarantee respect for privacy, thus strengthening efforts to prevent diseases and improve health security.

*the challenges of knowledge [in French]*, ADSP N° 93, December 2015.

# ARTICLE III
# DEVELOPMENT OF THE TECHNOLOGICAL LANDSCAPE AS A KEY SUCCESS FACTOR IN 'MAKING THE DATA SPEAK'

## 1. INTRODUCTION

### Why is IS architecture providing new opportunities for 'Mégadonnées[8]'?

The technologies specific to Big Data do not, in and of themselves, provide anything new in terms of the concepts for processing data themselves. However, they do in terms of the way that they are integrated in order to, on the one hand, meet the specific needs of 'Volume, Velocity, Variety' and, on the other hand, offer new opportunities in terms of 'Variability, Veracity, Value, Visualisation[9]'.

In dealing with Big Data, traditional technologies are limited by their rigidity in terms of data format ('Variety'), and by the non-scalability of their tools (Volume), the vast number and growth of available data sources and the growing number of analysis techniques necessary ('Velocity').

Big Data is ultimately more an evolution than a revolution. Today's Big Data will be tomorrow's 'Small Data'. The fact remains that the core is still 'the data'.

If we need to have a definition, the Big Data technological landscape designates all of the **technologies**, **architectures, infrastructures and procedures** that enable the very rapid **capture**, **processing** and **analysis** of large quantities of data and changing, heterogeneous content

8. French term recommended by the Official Gazette of 22 August 2014.

9. These requirements are often described as the 'Seven Vs' or the '7 Vs'.

in order to extract the pertinent information, on an industrial scale.

Focused on 'mégadonnées' (Big Data), this technological landscape must be able to evolve in order to adapt smoothly and quickly:

• to the availability of new data sources: compatibility of internal and external systems (cloud), migration, Internet of Things (IoT), etc.;

• to different formats of data, models, ontologies and semantics: structured and unstructured data, regulatory models;

• to the needs for adaptation of data flows: adjustment to operational or decision-making processes, etc.;

• to heterogeneous processing of data: storage capacity, storage area, aggregation, query, traceability, etc.;

• to varied analysis requirements: depth of analysis over several years, alignment of various data domains in order to respond to new questions, etc.;

• to changes in operational contexts and regulatory requirements: adaptation of models to market changes, predictive models, etc.;

• to the needs for personalised visualisation and communication: data adapted to the professional context and user profile (patient, regulatory authority, healthcare personnel, etc.);

• etc.

This landscape is constituted by additional technological components (bricks) that each meet a technical or functional requirement. Through the integration of these components, they represent a technological landscape that achieves the capacity and processing performance levels necessary for Big Data. The 7 Vs[10] therefore define the horizon of this technological landscape.

10. The 7 Vs are 'Volume, Velocity Variety, Variability, Veracity, Value and Visualisation'.

## How should the design of IS take into account the functions of knowledge generation?

**To define a landscape such as this, the principal subjects to be considered are the following:**

• the availability and granularity of the data essential for the scope of application sought;

• the data formats to be integrated/displayed;

• the appropriate volumes to be centralised, the acceptable confidence levels for the data to be processed, the requirements for freshness of these data, and the depth of the data (history);

• the types of processing to be implemented, the need to develop the algorithms through learning;

• the execution speeds appropriate for the processing operations;

• the flow rates for information to be considered;

• the ability to consume these data on a self-service basis;

• the need for space for prototyping and assessment of new scenarios in agile mode, of the 'Sandbox' or 'Data Lab' type;

• the variety of different tools for Data Mining, retrieval and display;

• the service levels expected by the various consumers of these data;

• the management strategy adopted for the processing of these data (duplicated, isolated or centralised);

• the good practices made available to meet these requirements (professional data model, etc.).

**And for each of these subjects, we must ask the following questions:**

• For whom?

• For what 'professional' objective?

## **A.** VARIETY

### What the technology provides …

**The richness of unstructured data can finally be exploited.**

One of the benefits of Big Data is the analysis of unstructured data (such as video, voice recordings, non-digitised texts, etc.), because production of these data is more abundant and their use provides greater added value.

In this context, it is not the '**Big**' that is important but rather the '**Smart**'. For example, to understand a document drafted in natural language, the choice will be based on the technologies and standards of the '**Semantic Web**' to perform semantic analysis.

### … and what this could mean for prevention

Scientists can now work on new data, irrespective of their original format.

As a result, the Mayo Clinic has introduced textual analysis tools in order to have access to a clearer understanding of the free-text notes from millions of medical records available in its systems. Six types of HL7 messages are now collected in near-real-time, indexed, transformed, analysed and made available to personnel.

## **B.** VOLUME

### What the technology provides …

**Volume allows exhaustive information.**

In 2019, the storage potential of all computers in the world should be 20 exabytes, namely $10^{18}$ bytes.

Sébastien Verger, Technical Director at EMC, predicted in 2013 that '*the data volumes will increase by a factor of thirty by 2020, reaching 35 zettabytes (namely $10^{21}$ bytes) globally*.'

Scientists, who are used to the scarcity of information and limited samples, are scaling up: information is now 'ordinary' rather than 'rare'; 'x%' samples are now 'all', although this sometimes happens at the cost of the quality of the information.

## … and what this could mean for prevention

Exhaustive information means that we can be less exacting about the accuracy of the information. The data are often in disarray, of variable quality and taken from a vast number of sources. The information from unstructured data is often not very dense, but its value lies in its quantity.

Working with complete data, even if they are imperfect, makes it possible to gain objectivity in terms of the principle of causality, which is often a source of errors and incorrect interpretation.

Biology is a domain where a hypothesis is the basis for all reasoning. The protocol is always the same: an initial hypothesis is stated and its accuracy or inaccuracy is then verified.

By way of example, rather than stating a hypothesis on the basis of an observation of a protein, the use of 'Data' makes it possible to identify a probable trend on the basis of a mass of data on a range of different molecules. This is what the American Ernest Fraenkel, a biology researcher at MIT, is working on. He is attempting to construct a unique model that incorporates all types of data: '*My greatest innovation was to propose a holistic interpretation of the data.*' Ernest Fraenkel is therefore revolutionising biology codes with this new approach.

## What the technology provides …

**Associated with Big Data, Data Mining is becoming 'Big Data Mining'…**

The tools for data analysis (statistics and Data Mining) and text analysis (text-mining) are essential for the use of repositories of information.

Current Data Mining tools make it possible to analyse numerous data, but over a sample considered to be representative. These tools are limited by the associated processing time and/or processing capacity.

Through the combination of Data Mining tools and Big Data tools (immense storage capacity, speed of execution of processing), Big Data Mining makes it possible to process all of the data with processing times that have become acceptable and overlapping, and to move from the descriptive towards the predictive and the prescriptive.

**The ability to combine all of these data for greater transversality and agility…**

With Big Data technologies, data can be factored and aligned in an analytical ecosystem that offers a level of transversality, making it possible to respond with greater agility to the challenges facing the industry. The same data are then consumed, from distinct professional standpoints, by means of Data Mining and business intelligence tools, by different professional profiles.

## … and what this could mean for prevention

Scientists need to generate statistical processing very quickly and to keep a historical record of those operations.

Adverse effects associated with medicinal products cause 10,000 deaths in France each year. In order to create alert procedures based on past errors, the European PSIP Project (Patient Safety Through Intelligent Procedures in Medication) is proposing, *inter alia*, that these procedures be generated on the basis of an automated data search (Data Mining).

Let's consider this in terms of the technological landscape. Suppliers of MRIs have the capacity to ensure a service quality and an availability rate for their sensitive equipment that is substantially increasing, which will be of direct benefit for physicians and patients. GE Healthcare relies on advanced analysis functions and various predictive models applied to the data sent by its equipment in real time to perform predictive maintenance. Intervention therefore takes place upstream, before any service interruption can occur (source: TeraData).

## **D.** THE 'VOLUME / VELOCITY / VALUE' RATIO

### What the technology provides …

**A considerable increase in the volumes processed without any loss of performance**

The data exchanged over the internet, originating in some cases from connected tools, are stored on servers and hard drives. They can be collected directly. Scalability is the ability of a system to maintain its functionalities and performance in the event of a significant increase in workload. But scalability models are not linear and the mere fact of adding storage capacity does not necessarily improve performance.

The traditional solutions available on the market (IBM, EMC, etc.) provide the 3 Vs required for Big Data, but each has its own specific model for the implementation of distributed storage: Cluster File System or Parallel File System. But these solutions do not provide the

same performance or the same degree of scalability in the event of a ramping-up of operations (when the storage capacity of the disks increases by 100,000, throughput only increases by 100).

As a result, the 'traditional' decision-making architecture with its database is no longer the only reference architecture. It is by thinking outside the constraints of traditional architectures that the players involved (primarily the big actors of the internet) have managed to find solutions. There are currently three complementary reference architectures to be managed: database, In-Memory and massively parallel.

### … and what this could mean for prevention

Having a website that is always available, being able to collect colossal quantities of information very quickly, dealing with peaks of flows to that site, ensuring the security and backing-up of the platform – these are necessities for any website of a health player wishing to disseminate information to or collect it from medical personnel and/or patients, while ensuring that the flow of information is secure.

Johnson & Johnson, one of the primary manufacturers of hip and knee implants, will be working with Watson to create a concierge service for patients in order to help them prepare better for surgery and support them during the post-operative period.

The latest generation of pacemakers are smart devices: they send real-time information to the hospital or doctor, making it possible to respond quickly in the case of an anomaly or at least to provide remote monitoring of the patient's medical status.

## **E.** VARIABILITY AND VERACITY

### What the technology provides …

#### The algorithms for 'what'

One of the principles of Big Data is not just to discover correlations

and trends but to do this without explaining the origin of those phenomena. A correlation is merely the quantification of the statistical relationship between two values. It will be a strong correlation if one of the values is very likely to change when the other is modified. Correlations do not provide certainty, merely probabilities. They do not say why something has occurred, but simply that it has occurred.

Big Data tackles the 'what' and not the 'why'. However, it does so with huge precision. There is therefore no longer any need to state hypotheses to be verified: we merely need to let the data speak and observe the connections between them. In some cases, we will not even be aware that these connections exist.

However, according to Frank Pasquale from Yale University (*The Black Box Society*), neutrality is a myth and the manipulation of algorithms is a reality.

Based on this observation, the technologies of Big Data are intended to push this neutrality forward, because they apply to a much broader scope of information, and therefore to provide greater objectivity.

## … and what this could mean for prevention

Last December, Google announced the creation of Verily, a new subsidiary bringing together its projects in the medical sphere and based on the former entity Google Life Science. Equipped with hardware, software, clinical and scientific teams, Verily works on platforms, products and algorithms intended to identify the root causes of diseases and to analyse the most appropriate treatments in order to provide better diagnosis and care.

We must, however, be careful to ensure that we always correctly qualify, contextualise and provide perspective for the data, because the degree of responsibility in the field of health is high.

In 2009, in the midst of the H1N1 flu pandemic, the American Ministry of Health asked Google for help. By locating the map position and origin of keywords typed in the well-known search engine, the engineers practically succeeded in tracing the course of the

epidemic, with a few minor deviations: the word 'flu' in the search engine reflected a simple concern on the part of the internet community about the approach of winter.

# **F.** VISUALISATION

## What the technology provides …

### Visualisation for interactions

Big Data technologies make it possible to consume information completely independently, to have access to maximum interactivity and to obtain instantaneous responses.

Visualisation can make it possible to detect invisible trends immediately, and to highlight the most convincing data.

## … and what this could mean for prevention

Visualisation is not just a method of representation, but also a means of encouraging the stakeholders to accept the data, study them, share them and better understand them. A tool to support understanding and decision-making, it can become a method of communication.

This complex problem of data visualisation is the core of a programme designed to explore the culture of health driven by an American foundation. For example, which method of visualisation is the most appropriate to communicate about the importance of preventing cardiovascular disease? These graphic designers have no understanding of prevention and have thus worked on a list of 16 scenarios for the communication of medical risks. At the end of the chain, citizens and patients judged the various proposals.

## G. CURRENT DIFFICULTIES

**In an technological context that is still relatively immature:**

- difficulty in expressing the requester's requirements to the provider;
- difficulty for the provider in understanding the requester's medium- and long-term roadmap;
- difficulty for the requester in identifying the first structuring project and the successive steps leading to the medium- or long-term roadmap;
- underestimation of the efforts required to align and clean up the data;
- multitude of technologies and technological combinations;
- emerging expertise of integrators;
- absence of objectivity about the industrial capabilities of open-source platforms in terms of serving operational requirements and professional requirements;
- underestimation of the requirements for data governance;
- difficulty in obtaining a sponsor for general management that is able to drive a transverse project of this kind.

## H. CONCLUSION

- The functional possibilities provided by Big Data are immense.
- The regulations are still open to proposals for technological innovation.
- The principal challenges are still organisational or ethical.

## I. A UNIFYING INFRASTRUCTURE – ESSENTIAL FOR BIG DATA

### What the technology provides …

Managing such a rich and complex environment requires a new

approach by the technological landscape.

Therefore, we talk of an 'ecosystem' that is able to process the data effectively in all of its expressions and dimensions, to support the use of innovative tools for visualisation and analysis, to optimise the data integration process and to adapt immediately and without compromise to the various constraints (technical, operational and regulatory).

The parallel and close-in processing of data (In-Database) are essential and represent a vital characteristic of the Big Data infrastructure. Processing operations are performed in parallel, a unique guarantee of response time and service level commitment that is essential for serving users whose requirements increase with the expansion of the data available.

The intrinsic industrial capabilities and flexibility of a platform of this kind therefore enable the combined use of a range of data processing technologies. This modular, agile and coherent system is described perfectly by analysts such as Gartner using the term 'Logical Data Warehouse Architecture Model'.

This architecture model allows organisations to access the data, refine them and retrieve them in a way that is reliable while at the same time guaranteeing the availability of the system, and the performance, concurrency and variety of processing operations.

Selecting the appropriate tool to address the problem identified is a key principle of good engineering practice. On this fundamental principle, the concept of a Big Data ecosystem is intended to provide the right technology to perform the right processing. The adage '*a good workman chooses the right tools for the job*' applies perfectly. There is not currently one technology that is capable of resolving all problems.

This agile architecture can be broken down into the following five components:

1) the Data Lake;

2) Data Discovery;

3) an Integrated Data Warehouse;

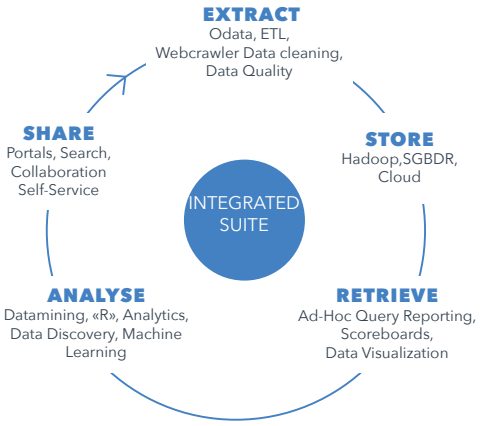
4) the Data Integration Layer;

5) a proven resource allocation engine (Integrated Workload Management).

This architecture is fully integrated and interoperable.

The key is to understand which parts of an analytical workload should be allocated to each of the technologies within the architecture and to orchestrate the interaction of these various components. The analytical workload can involve several sources and types of data, requiring a dispersal of processing operations over different technologies within the ecosystem.

For example, all of the components of this architecture will be needed for the correlation of behavioural data originating from the mobile application providing the monitoring of patients with type 2 diabetes (activities, recording of vital signs, glucose level, well-being assessment, meals and calories, etc.), with data from the community website, prescription data for a patient population and the individual medical file. The Data Lake will be needed more for the integration of data from smartphones or the community website, while Data Discovery will rely on the Data Integration Layer to ensure transparent access to the data in the Data Lake and the Integrated Data Warehouse. Data Discovery will then apply the various algorithms for the analysis of behaviours, and lastly the Integrated Data Warehouse will be used for more structured data based on prescription or the patient file, or for example the recovery of the results from the Discovery phase in order to operationalise these results vis-à-vis patients.




*The various stages of Big Data processing. Source: CXP 2015*

### Data Lake

The Data Lake component houses the raw material: unrefined raw data with a limited information density, entered without a strict organisational structure so that it can then be transformed and correlated in order to provide new perspectives and professional value.

The continuous capture of all data is an implicit goal to be achieved, because any data ignored and therefore not used will never be able to reveal their potential. The implementation of a successful Data Lake must meet the following key characteristics:

- scalability;
- low cost of the stored data per terabyte;
- support for structured, semi-structured and unstructured data;
- openness to external solutions and tools.

### Data Discovery

The standard use for the Data Discovery function falls within R&D activities undertaken by data scientists and business analysts as part of an analytical approach based on exploration and statistical modelling. It is unique in its enormous agility to correlate data originating from the Data Lake, Integrated Data Warehouse or external sources, and in its quick iteration in order to find and rapidly reveal hidden facts and proven correlations. It also offers business users the option of using these results in the form of reusable applications.

The implementation of a successful Data Discovery solution must meet the following key characteristics:

- The vast range of available analytical techniques (statistical and textual analyses, analysis of graphs, path analysis, analysis of models, etc.) and script execution techniques (R, SAS, SQL, MapReduce, etc.).
- The effectiveness of the process of data acquisition, preparation, analysis and visualisation.
- Support for structured, semi-structured and unstructured data;

• Openness to external solutions and tools.

### Integrated Data Warehouse

The Integrated Data Warehouse component integrates and deploys the data as a finished product for business users for operational, decision-making purposes. Its role is to unify and bring together a single version of the data, to be used for multiple purposes (multi-domain management) to meet varied user requirements. The degree of transversality sought will be all the more effective where the information system is actually integrated as a whole.

The requirements associated with multi-domain management mean that the solution must be able to provide constant performance demonstrated by contractual service level commitments vis-à-vis end users.

The solution must also make it possible to mix production data with non-certified data within analytical Data Labs or Sandbox. In other words, self-service Data Labs enabling users to create a space for prototyping and evaluation in which they can test, verify and validate hypotheses by combining certified production data (read-only) with new sources of information (flat file, real-world data, data from the Data Lake or Data Discovery).

The implementation of a successful Integrated Data Warehouse must meet the following key characteristics:

• a high-velocity solution built on massively parallel processing technology;

• robustness and high availability of the solution;

• linear scalability in line with the change in requirements and constraints;

• analytical functions performed as close to the data as possible (in-Database);

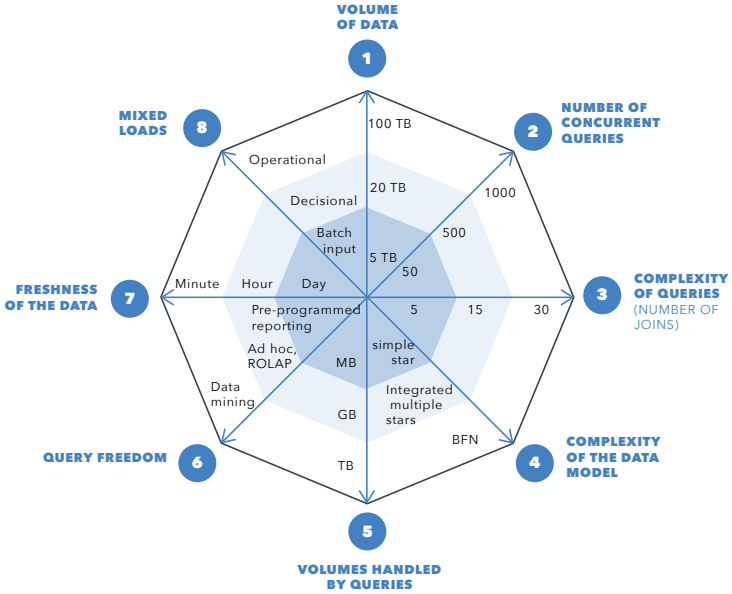
• support for structured, semi-structured and unstructured data;

• advanced logging, traceability, auditability and security functions;

• flexibility of self-service spaces (Data Lab or Sandbox);

• openness to external solutions and tools;

• the ability of the solution to absorb operational constraints simultaneously (volume, number of simultaneous queries, performance).




*Simultaneous development over several dimensions*

### Integration Data Layer

To be productive, users do however need technologies that free them of the need to transfer data from system to system, or even to understand the mechanisms by which data are moved.

This component is particularly important at a time when open-source technologies are becoming a major component of Big Data environments.

It is true that different technologies originating from different suppliers are being used together within analytical and operational systems. These non-integrated technologies are often considered by users to be distinct systems. It is therefore difficult to exchange data and to perform processing operations between separate components.

The Integration Data Layer provides a layer of interoperability and therefore transparent two-way interaction for data and for analytical processing operations between the various components of the Big Data ecosystem. It takes care of managing the performance of queries that use several analytical engines and data warehouses (Data Lake, Data Warehouse, etc.), in a manner that is optimised and secure, thus integrating multiple systems so that they create a single unified system.

The implementation of a successful Integrated Data Layer must meet the following key characteristics:

- two-way processing;
- optimisation function during the performance of queries on multiple systems and heterogeneous technological environments;
- integration with open-source environments;
- possibility to query the ecosystem from any component (Data Lake to Integrated Data Warehouse, Data Discovery to Data Lake, Integrated Data Warehouse to Data Lake, etc.) ;
- limited data movements or data duplication.

### Integrated Workload Management

Effective workload management is crucial in providing corporate users with responses to their questions in real time and meeting performance and service level commitments for Big Data infrastructures. The Integrated Workload Management component deals with dynamic resource allocation. Its role is to simplify and automate data management, optimise the allocation of machine resources and manage the coexistence of very different uses on the same platform. This means that each task can be performed under optimal conditions, guaranteeing the various service levels in terms of operational variations.

Through 'deterministic prioritisation', the engine ensures that the most important tasks are performed first. Its mechanisms also make it possible to stabilise response times and to optimise the use of resources during peaks, whether planned or unplanned.

The implementation of a successful Integrated Workload Management system must meet the following key characteristics:

- holistic vision of the Big Data ecosystem (Data Lake, Data Discovery and Integrated Data Warehouse);
- dynamic changing of priorities without halting queries or the system;
- maturity of the dynamic resource allocation engine;
- hierarchised task allocation (priorities);
- filter, exception management and planning functions;
- definition of allocation criteria (rules);
- possibility of allocation of service levels;
- integration into an integrated surveillance tool.

### … and what this could mean for prevention

The GLIDE project (Global Integrated Drug Development Environment), implemented in 2015 within the Roche pharmaceutical group, combines a vast range of internal and external data within a single ecosystem (clinical studies/new processing data – generally SAS Datasets, laboratory data – blood, x-rays, electroencephalograms (EEG); medical data, genetic data, biomarkers or even real-world data).

This analytical ecosystem has the effect of significantly reducing processing time (from several days to some hours or minutes), but also of enabling R&D teams to develop better targeted therapies, reduce the adverse effects of new molecules or even, using all of this historical data of observations, reposition drugs for new therapeutic purposes.

## J. GLOSSARY

**Unstructured data:** texts originating from the processing of texts or content of electronic messages, audio files, voice files, fixed images or videos.

**Structured data:** spreadsheet pages organised in tables, databases where the data are organised in connected tables.

**Big Data:** literally 'big data'. English expression describing all data that are so big that they become difficult to process using traditional database management tools. Within the HDI, the term has a broader meaning and describes all data irrespective of source, format and volume.

**Smart Data:** focuses on the data relevant in relation to their specific objectives.

**Data Mining:** detection of information in a database.

Tools capable of detecting the information hidden 'at the deepest levels' of the 'data mine'. This does not concern database query systems, spreadsheets, statistical systems or even traditional data analysis systems.

Data Mining is undertaken using several different approaches:

• **'Verification' approach:** the user has an impression or general idea of the type of information that can be obtained from the data. The database can then be used to 'quantify' that impression. It is clear that the data extracted, and the decisions made using those data, depend exclusively on the user's impression of the important parameters of the problem (age, geography, etc.), an impression that is often correct but not exhaustive.

• **'Discovery' approach (Advanced Data Mining) or search for hidden information:** the user understands that the quantity of data to which he or she has access is considerable and thus the optimal and exhaustive detection of important structures or relationships is completely beyond the reach of human users. Users must therefore use advanced data analysis methods to detect the hidden information (which may be the most interesting).

**Text-mining:** all of the methods, techniques and tools for making use of unstructured documents. Text-mining relies on linguistic analysis techniques.

**Architecture:** describes the general structure inherent in an IT system, the organisation of the various elements of the system (software

and/or hardware and/or human and/or information) and the relationships between them.

The following list includes some different architectures:

• **Traditional decision-making architecture (Oracle, SQL Server, MySQL, Informatica, Datastage,** etc.): this architecture performs well where the volume of data to be transferred between each stage is limited.

• **In-Memory architecture (Qlikview, ActivePivot, HANA, etc.):** this architecture can offer high-performance analysis services, or indeed real-time services (updates to data and recalculation in line with clusters) and simulation services.

• **Massively parallel architecture (Hadoop, TeraData):** this architecture can store an immense quantity of data (unlimited) elastically.

**Scalability:** ability of a system to function correctly with increasing workloads.

## **2.** 'KNOWLEDGE BY DESIGN' OR SEMANTICS AS THE KEY TO VALUE

The information systems within the health sector should in all cases be able to contribute to public health, to support our health system and to generate new knowledge. This is only possible if we define a certain number of principles and rules that create a coherent, stable and multidimensional technological, legal, financial, organisational, institutional and political framework. Within this 'planning and development' approach, the production of data, their type and the conditions under which they can be reused represent a major investment for the future. The creation of value expected as a result of the digitisation of the sector and the large-scale production of data therefore require a strong commitment in terms of interoperability.

### Semantics, a source of infinite value creation: 'knowledge by design'

The deployment of internet infrastructures and the organisation of the various communication protocols benefit all business sectors. The specific issues affecting the health sector cannot be found here, but rather in the more 'business-specific' layers, first among which is semantic interoperability, the principal source of value with the emergence of the Internet of Things. And for health, this constitutes a complex lever to be grasped and understood, since it carries enormous promise. Producing electronic information certainly provides direct savings in terms of time and money. But ensuring that these data make sense and can be interpreted, analysed and reused creates opportunities that are simply infinite. Semantic repositories are the printing press of the 21st century and will make it possible to produce and transmit the bulk of knowledge over the coming centuries. Data production is experiencing considerable growth and this production must therefore have access to such repositories as quickly as possible, because we will not be able to turn back the clock on those data. The challenge of being able to automate some or all of the processing of these data taken from multiple sources and produced in contexts and for purposes that are different in each case is becoming an extremely urgent imperative. The deployment of an IT structure that continues, as is still often the case today, to use electronic data that are unstructured and not reusable must be replaced as soon as possible by the dissemination of semantic repositories instead of, for example, unusable JPEG documents. In more macroeconomic terms, the capacity to reuse data will help to share the costs of production and enhance the creation of value. The right to reuse data for purposes referred to as 'compatible' introduced by the new European Regulation is a measure that has substantial impact. The model that until now required that data be generated specifically for a dedicated and most often unique purpose should now be reserved for ad hoc studies making use of concepts that are insufficiently widespread to enable the use of a common vocabulary.

The semantic integration of different systems constitutes a major challenge that must be taken into account far upstream, in order to enable the application of a 'translational' approach establishing a

permanent bridge between treatment activities on the one hand and research on the other. This concept of knowledge by design, by analogy with privacy by design, promoted in the field of security, must ensure that the bulk of the data produced has a common meaning that can benefit public health in particular.

Other countries are concerned with this issue, as we can see from the creation of the CDISC consortium (Clinical Data Interchange Standards Consortium) by the American Food and Drug Administration. This agency has the task of sharing clinical trial data around common, free-access semantic repositories. We should also consider the strategic linguistic issue. The 70 countries making up the French-speaking world and the 220 million French speakers worldwide have an interest in making sure that English-speaking codifications are not the only such systems in general use.

Each item of health data can potentially contribute to prevention, provided that its meaning is clear and that the reuse of the data is taken into consideration when the system is designed.

# ARTICLE IV
# THE DEFINITION OF A NEW PARADIGM FOR STAKEHOLDERS

## 1. BIG DATA: A NEW PARADIGM FOR THE PHARMACEUTICAL INDUSTRY

The pharmaceutical industry understands today that the quantity of available health data, combined with current analysis capacities and algorithms, constitutes a major factor in the emergence of new practices in epidemiology, personalised medicine and prevention, and for research and development of new health services that will transform patient care and management.

A large number of pharmaceutical laboratories are developing medical products intended for patients suffering from chronic diseases, which require long-term monitoring. A drug alone is no longer seen as an isolated treatment variable. It forms part of a broader treatment scenario that has a decisive impact on long-term treatment compliance by the patient.

The drug is now just one element among others making up integrated health solutions that form part of the 'beyond the pill' strategy.

To ensure the efficacy of a drug, the pharmaceutical industry needs to have access to tools to measure the results: improvement in clinical results, and the prevention of the adverse effects and serious events that may occur in patients. In a patient-centred model, statistics play a vital role. They contribute to understanding the behaviour of the patient and thus make it possible to identify the factors that determine treatment compliance by the patient[11].

Pharmaceutical laboratories are moving in this direction by developing and providing innovative health services for patients: not just

11. Interview with Pascale Witz, *Healthcare Data Institute Newsletter*, June 2015.

medicinal products, but also the accompanying healthcare solutions. Many of these solutions involve the collection and analysis of data in real time using appropriate algorithms, so that quick treatment decisions can be made.

The technology will change the patient experience (by improving care and cooperation by the patient) and, by enabling the analysis of large quantities of data, will make it possible to initiate a new management and care process. Improved clinical results for patients is the objective to be achieved in order to consolidate a suitable business model. As a consequence, the technology combined with therapeutic monitoring will have greater impact compared to technology alone (see Diagram 1).


*Diagram 1: The technology of connected objects is changing the patient experience, and the collection and analysis of Big Data are facilitating the process. The diagram shows an example of an integrated treatment approach for monitoring and supporting patients suffering from respiratory problems.*

Several innovative software systems for health data have already moved on from the stage of retroactive reporting (data collection) to incorporate predictive functionalities (data analysis), making it possible to warn patients of a potential adverse effect.

At the same time, recent technological advances have facilitated the collection and analysis of information originating from multiple sources; a major advantage in the area of health, given that the

personal data relating to a patient can come from various measurement technologies, in many cases with little or no active participation from the patient. For example, portable ingestible or implantable sensors and portable digital applications are able to collect and transmit data independently; the analysis of data essentially in real-time will enable healthcare professionals to anticipate and avoid imminent crises in patients.

## **A.** PATIENT MONITORING AND INDIVIDUAL SECONDARY PREVENTION – CHRONIC DISEASES AND DIGITAL HEALTH

The value of prevention covers both prospective prevention of populations and individual secondary prevention, the latter being specifically intended to avoid or anticipate serious events, relapses and crises in patients suffering from chronic diseases (such as diabetes, asthma, rheumatoid arthritis and cardiovascular disease).

The implementation of individual secondary prevention generally requires integrated health solutions that include medical devices, portable sensors, applications for smartphones, the collection and analysis of health data in real time using appropriate algorithms, and real-time feedback and notification of patients and healthcare professionals, in order to enable interventions in emergency cases (see Diagram 2).

Intelligent devices (such as wireless blood pressure monitoring, devices used for passive and active data collection, wristband sensors for monitoring sleep quality) are the common components of advanced platforms for patient surveillance and participation, which are designed to improve patient treatment and management, and therefore to achieve better clinical results and reduce the total cost of treatment.

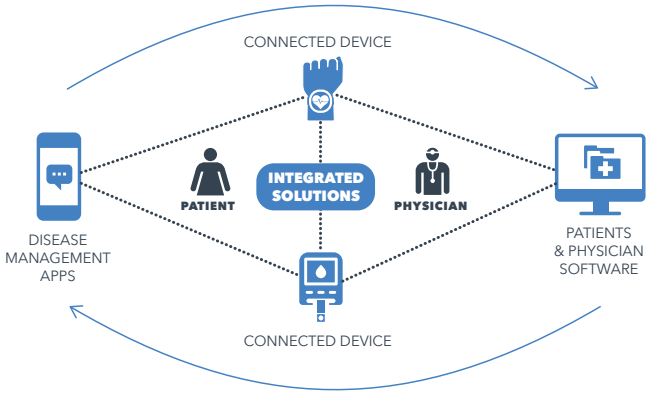



*Diagram 2: Connected objects, sensors, software and data analysis are the necessary components of integrated health solutions.*

Doctors and patients are beginning to use digital tools that enable more effective management of health. In the future, pharmaceutical laboratories and other stakeholders will adopt the existing technologies and develop new ones. As part of that process, they will create a *more connected digital health ecosystem* that will benefit everyone:

- **The pharmaceutical laboratories** are creating applications and other platforms to enable better management of diseases and treatments, providing information about illnesses, drugs, appropriate doses, expected clinical results, health objectives and healthy life choices.

- **When a doctor** prescribes a drug, he or she will also prescribe the digital application.

- **Patients** can also collect and monitor their own data in real time. They can send them to their GPs for in-depth assessment, and make adjustments themselves to their drug treatments or lifestyles.

- **Patients** can also send their data to pharmaceutical laboratories, which may then use this information to develop treatments and improved health services targeted more on patients, and thus conduct corresponding research.

## **B.** EXAMPLES OF INITIATIVES FOR INTEGRATED HEALTH SOLUTIONS USING REAL-TIME DATA ANALYSIS

*Ginger.io* is offering a mobile application in which patients suffering from specific diseases can ask, in cooperation with doctors, to be monitored by means of their mobile telephones and supported in the form of behavioural health treatments. The application records the data on calls, text messages, geographical location, and even physical activity. Patients also respond to questionnaires provided on their mobile phones. The *Ginger.io* application integrates patient data with public research on behavioural health by the National Institutes of Health and other sources. The information obtained can show, for example, a lack of movement or other activity that would indicate that the patient does not feel physically well, and that there are irregular sleep cycles (revealed by late calls or text messages) that could indicate that an anxiety attack is imminent[12].

### Diabetes

There is no absolute rule as to the dose of insulin to be injected. When a patient begins treatment for diabetes, he or she must move through steps (dose-ranging). The challenge for diabetic patients is to maintain the levels of sugar in the blood. Hypoglycaemia (underdosing) or its opposite hyperglycaemia (overdosing) can be the source of serious problems.

A solid treatment plan will include a solution for the continuous monitoring of glucose levels, based on data from connected medical devices and analytical data capabilities, in order to prevent a hypoglycaemic episode well before the clinical symptoms manifest. The *Diabeo* solution from Sanofi is one example of an advanced technology providing a telemedicine solution enabling the individualised adjustment of insulin dose, based on glycaemic measurements, patient treatment regimes and physical activity.

12. *Unlocking the full potential of data analytics for the benefit of all,* Healthcare Data Institute 2015.

Mobile monitoring of vital signs for diabetic patients using portable (wristband) connected objects makes it possible to share data in almost real time with the patient's healthcare professionals, and thus anticipate the risk of future complications, including diabetic foot and ocular conditions.

An automatic portable bionic pancreas using a sensor connected to a smartphone and with algorithmic software has been shown to perform better than traditional insulin pumps, according to an article published in the *New England Journal of Medicine*, July 2014.

### Interesting White Papers

- *Predictive Medicine Depends on Analytics*, Harvard Business School
- *The Big Data revolution in healthcare*, McKinsey & Company

## **2.** THE INSURANCE SECTOR AND BIG DATA: PREDICTIVE AND THEREFORE PERSONALISED HEALTHCARE?

### **A.** BACKGROUND

We live in an era of revolution and technological change defined as the third industrial revolution. The source of this revolution is the datum, by analogy sometimes compared to what was represented previously by coal, and then oil, in their transformations leading to the mass production that is always at the heart of our current society.

Crude oil, like raw data, is nothing in itself. Refined and transported, it created the petrochemical revolution and the current transport revolution. Placed correctly in context and analysed using algorithms that benefit from the constant increase in growing computing power, raw data will lead to an incredible creation of value.

This revolution is based on the capacities of Big Data. These capacities are based first and foremost on an exponential increase in the volume (V) of data available due to the increase (**V for Variety**) in

connected objects (smartphones, PCs, watches, consoles) and sensors (altimetry, home automation, vehicles, household appliances, etc.). All of these increasing interconnections define the concept of the Internet of Things (IoT) or the Internet of Everything (IoE), which generates a doubling of the volume (**V for Volume**) of the data available every three years. Some people are calling this Internet of Things the new steam engine of the 21st century.

The consequence of this 'breaking wave' is the 'datafication' of the world or the digitisation of the real. We should note that the analogy with coal or oil, as finite resources, stops there. Data are an inexhaustible resource.

Big Data is also based on the current power (**V for Velocity**) of storage capacity and calculation capacity, and on the sophistication of the algorithms.

For the first time in the history of humanity, we have access to a volume of data and a computer power that makes it possible to escape the constraints of sampling, accuracy of measurement or the need to have to work on limited data analysed using standard statistical calculation.

While this data revolution will greatly affect all economic organisations and sectors, **three sectors are particularly impacted: Housing** (home automation, energy saving, household appliances), **mobility** (connected and independent vehicles) **and health** (efficiency of care, ageing, monitoring of chronic diseases).

These sectors are the primary markets for insurers and they play a major role in them.

Where these markets meet, **insurers will therefore be compelled,** by the Internet of Things and Big Data**, to rethink their service offers, internal organisations and business models**. This will mean, on the one hand, that they will not be left behind and, on the other hand, that they will be in a position to exploit the extraordinary opportunity to create value (the fourth V of Big Data after Volume, Variety and Velocity) presented to them.

# **B.** IMPACTS AND PERSPECTIVES

Insurers are historically collectors of data which they then use for analysis. These analyses are based on claims made, on a vast set of societal data and on a pooling of risk enabling the calculation of premiums.

**Until now, the analysis and assessment of risk have paid little regard to the use made of the insured item,** whether this was tangible (property and casualty insurance) or intangible such as health (health insurance). **The new techniques** for the collection of data, the more widespread use of connected sensors, the digitisation of health with its increasing flow of associated data and the introduction of ad hoc algorithms able to adapt in order to follow the changes in the behaviour of an insured person **will shatter the traditional model of actuarial calculation.**

For insurers, **the issue is to bring together the aggregated historical knowledge they have of their insured people and claims** (offline data) **with the data from the new digital environment** (online data) that are now accessible (visiting of websites, activity on social networks, home automation, on-board units, connected devices, etc.). This merger of the offline and online enables targeting by the analysis of weak signals and predictive, personalised monitoring of insured persons.

This new use of data impacts the entire insurance value chain, namely:

- design of products and pricing;
- knowledge of the customer and the market;
- prevention of fraud;
- management of claims;
- new services.

One preliminary statement should be noted: in France, the storage and use of these personal data are governed by very strict regulatory requirements, particularly in relation to health data, which

significantly restricts the possible avenues open to insurers.

## The design of products, pricing, knowledge of the customer and the markets

For the design of the products, and to refine understanding of the customers and the market, marketing has or will have access to increasingly high-performance tools intended to improve segmentation, personalisation and identification of insured persons.

As explained above, the increasing number of sensors, the reconciliation and cross-referencing of data, and the power of the algorithms mean that we can now no longer look directly and only at the value of the number of claims but also at the use and behaviours observed in relation to the scope insured.

There are countless examples and possibilities. Offers are already appearing that make it possible to differentiate certain behaviours of drivers in terms of premium. In the USA, certain health insurance negotiations depend on the results identified by connected objects.

Similarly, the analysis of weak signals from the IoT makes it possible, for example, to reduce the risk of attrition (loss of an insured person) that can occur during a change of vehicle (detection of unusual queries on websites with vehicle offers) or housing (surfing on property sites) and anticipate this loss by proposing tailored offers.

Pricing will benefit from all the power of the new tools. The usual variables (sex, age, type of vehicle or housing, etc.) are being enriched by behavioural and usage data that enable a segmentation of the product offer and real-time adaptation of premiums.

Insurers are moving from a descriptive view of the events on which actuarial calculations are based to a predictive view that makes it possible to adjust the management of risk as closely as possible to the person.

## Fraud prevention

This issue is vital for insurers. Close to 20% of fraud cases are not detected. The cost of these cases weighs heavily on the claims process and therefore on the calculation of premiums. Through its analysis of data and weak signals, and through its ability to detect unusual behaviours and inappropriate usages, Big Data promises a potential significant drop in this cost. Quite apart from an associated return on investment, the margin provided suggests there could be a fall in contributions and/or the design of new offers.

## Management of claims

In addition to the reduction in the number of claims expected as a result of the monitoring and personalised knowledge mentioned previously, the increasing digitisation of process and flows, and the resulting internal restructuring, will enable an optimisation of the management of claims, which adds to potential productivity gains.

We should note that with regard to France, the implementation of the health insurance card (Carte Vitale), combined with remote transmission from local health insurance funds (CPAM) to top-up health insurance agencies, has already broadly permitted this optimisation.

## New services

These new services are generated by the new marketing tools that identify new offers and requirements resulting from predictive, personalised monitoring. These new services are improving CRM, increasing customer loyalty, reducing attrition, and promoting a reduction in the cost of recruitment campaigns. They are also contributing significantly to a strategy of differentiation and the search for a competitive advantage.

In short, insurers can expect that the Internet of Things, connected tools and Big Data will provide them with increased productivity and improved operational quality. Above all, they can refine their product offers and their services for insured persons (gains in well-being, reduced or adjusted premiums, etc.).

In this regard, French legislation, which provides considerable protection for consumers (in particular the national data protection agency, the CNIL), could become an asset by making it possible to organise the collection and provision of all necessary data, **and establish a relationship of trust between insurers and their insured customers.**

These insured customers must retain control of their data and must be informed (information in customer spaces on the insurers' websites, communication push and targeted information campaigns) and convinced about the uses and purposes for which their data are collected.

Transparency makes it possible to establish trust, the trust to authorise collection. Collection makes it possible to achieve the critical volume that enables analysis.

Analyses lead to personalised or predictive diagnosis using algorithms that embody the expertise and industrial secrets of the insurer.

These algorithms will become the core of the value added by an insurer and the guarantee of its performance in the market, in strict compliance with the legislative context and an ethical approach that favours the individual in all situations.

## Insurance and health

In the health sphere, we should note that there is pressure on health insurers. In most developed nations, given the increasing scale of chronic diseases, healthcare expenditures are growing faster than GDP. In France, this cost is above 11% and we have seen this amount essentially triple over the last three decades.

This increase can be explained by the increasing age of the population (as there are therefore more chronic illnesses to be treated over longer periods), by the high cost of new treatments (new oncology molecules) and by the new techniques for examination and surgery (imaging, catheterisation, transplants).

The emergence in daily practice of nanotechnologies and uses of genome technology are unlikely to reverse this trend.

These continually growing costs are essentially consumed by curative

treatments and leave only a very limited amount for prevention (about 2%).

As a consequence, pressure is growing on the top-up portion of the market devolved to health insurers. This tension can also be seen in the areas of dependency, disability and ageing.

This trend is also apparent in countries where the role of private health insurance is more significant than it is in France. The problem in this situation will then be to keep the amount of the premium at an acceptable level for the payer.

Health insurers are therefore seeking new business models that can address these difficulties. In this context, the Internet of Things and Big Data, because of their capacities in relation to prediction and individualisation of risk, are seen by the sector as new tools that can generate value while adhering to the traditional pooling model. Indeed, more specific pricing of the risk for an individual within a group does not prevent the pooling of risk (within a company, a field, a sector or even on the basis of principles associated with non-discrimination), and, conversely, new, more efficient forms of pooling should appear.

Furthermore, **top-up insurers,** through their role as financiers, must be able **to become more important players in the patient/care system interaction,** as a better understanding of insured individuals will be reflected by the provision of services or options that are even better suited to the needs of those individuals.

It is therefore necessary to acknowledge that, as part of the digital economy, the progressive and growing introduction of ICT and NBIC will result in a very deep-seated disruption to this interaction: movement of treatment paradigms towards prevention and well-being, changes in the way the health system is organised, new behaviour of patients (sometimes identified as consumer-players) who are involved and connected (social networks, websites, etc.), modification of the patient-healthcare professional relationship, etc.

In summary, the digitisation and 'datafication' of health are creating new forms:

• of the coordination of patient monitoring (platform of medical and welfare services, websites, etc.);

• cooperation between healthcare professionals by means of in-depth interaction and permanent connection;

• management by the patient, including when mobile, of associated health information (storage, sharing, summary, monitoring, alert, transmission, etc.);

• sharing and distribution of knowledge between players, including insurers and any new entrants (GAFA).

These changes are or will be reflected by:

**• the optimisation of the care path;**

**• a reduction in hospitalisations;**

**• an improvement in treatment compliance;**

**• a better allocation of resources intended for prevention.**

Health insurers, as financiers of risk and financiers of the system, can only become involved and support these trends driven by the Internet of Things and the tools of Big Data.

At the same time, the concept of health and the way in which it is managed are changing and expanding. Indeed, on the basis of the well-known definition from the WHO (a state of complete physical, mental, and social well-being, etc.), a change is taking place, with a shift from the the historical purely biomedical concept towards a holistic approach that attempts to understand all of the needs of a person (affective, health, nutritional, cultural, etc.). The determinants for health are becoming broad and multiple (behaviour, environment, etc.).

**We must therefore learn to manage health as a rare resource on the basis of this holistic approach, which will give priority to prevention rather than cure**.

The promotion of health thus put in place fosters the development of prevention and, in particular, the notion of well-being. It is in these two sectors (prevention and well-being) that the connected sensors and Big Data analyses are particularly pertinent and powerful.

This avenue (management of health capital to thus reduce health risk, extensive use of smart sensors and Big Data) is particularly identified and favoured by insurers, who are expecting positive impacts

throughout their value chain and namely the following:

• A possible financial saving (monitoring of chronic diseases as part of secondary or tertiary prevention, decrease in the number of claims through an improvement in hospitalisation days, improvements in drug compliance, a fall in the undue cost of avoidable services, etc.).

• A strengthening of the customer relationship (better known, and therefore better monitored).

• The possibility of new services (decrease in the rate of attrition, better customer loyalty, policy of differentiation, search for comparative advantages). This is one example: monitoring of admission/discharge from hospital, prevention of musculoskeletal problems, prevention of work-related stress, nutritional coaching or coaching in physical activity, etc.

• An adjustment to premiums, increased productivity, reflected in an improvement in operational quality.

The culmination of this trend initiated by smart tools and the 'datafication' of health will accelerate through the arrival of nanotechnologies but above all through the increasing role of genomics.

The analysis and decoding of the human genome are being automated to an increasing extent, and the significant cost reduction will enable greater use of this technology in treatment. The power of the technologies implemented by Big Data makes it possible to detect areas of high risk potential and to optimise the therapeutic and insurance response.

This genome revolution is completing a fundamental paradigm shift, namely **the switch from standardised medicine** (the same treatment for a given condition irrespective of the population affected) **towards personalised medicine** (one illness, one patient, one environment, one treatment), also referred to as precision medicine. It is in this new landscape that the insurance business of tomorrow is developing.

The question of trust, already mentioned above, is one that is becoming particularly sensitive. Of course insurers must comply with the laws and legislative frameworks relating to health data and the protection of privacy.

They must also demonstrate transparency and know how to convince insured individuals of the relevance of their approach and the positive benefit they can expect from the collection and analysis of all the data gathered.

Public authorities must also move our current legal system forward in order to protect both personal freedoms and patient rights and the competitiveness of French insurers compared to their international competitors.

# ARTICLE V
## ANONYMISED DATA, PSEUDONYMISED DATA: WHAT DEGREE OF AGGREGATION?

The terms 'anonymisation' and 'pseudonymisation' describe the processes now becoming widespread that are applied to health data, and make it possible either to sever the link between the identity of the person and the data about that person, or to be able to 'chain' the information about that individual without knowing his or her identity.

Anonymisation (or de-identification) of personal data describes the method and the result of processing personal data in order to irreversibly prevent the identification of the person concerned. In general, it is not enough to directly delete the elements that are themselves identifiers in order to guarantee that any identification of the person is no longer possible. An effective anonymisation solution must prevent reidentification, which is not limited simply to preventing identification (isolating an individual within a series of data, finding the name and/or address of a person) but also correlation (linking together distinct series of data about a single individual) and inference (deducing information about an individual from this series of data).

Pseudonymisation is a technique that involves replacing an attribute (generally a unique attribute) by another in a record. The result of pseudonymisation may be independent from the initial value (such as in the case of a random number generated by the data controller or a name chosen by the person concerned) or it may derive from the original values of an attribute or a series of attributes, for example using a hash function or an encryption system. The physical person is therefore always potentially able to be identified indirectly. As a result, pseudonymisation does not make it possible, by itself, to produce a series of anonymous data; it reduces the risk that a series of data will be correlated with the original identity of the person concerned. In this regard, it is a valuable security measure, because it reduces the risks for the person concerned, but it is not a method of anonymisation.

These definitions are in line with those stated in the Opinion of the

G29 Working Group of 10 April 2014[13] on the issue and of course with those laid down in the new European Regulation on the protection of personal data of 27 April 2016[14].

With regard to the protection of personal data, pseudonymised data remain personal data and are not intended to exclude any other data protection measure.

It is interesting to note that in the text of the European Regulation, which will soon become law in all Member States of the European Union in relation to the protection of personal data, pseudonymisation is also covered in Article 6(4) on the lawfulness of processing as one of the appropriate safeguards that the data controller may put in place to justify the possibility of processing data initially collected for a given purpose for another purpose considered as compatible.

This new notion of compatibility is important in legal terms to justify the further use of data for purposes different from those initially envisaged.

The exponential availability of data from different sources in relation to a single individual and the new capacities for processing these data, and in particular Data Mining, considerably alter the concept of reversibility of the anonymisation of personal data (reidentification). Data considered to be anonymous when they are provided can subsequently present a high risk of reidentification due to the introduction of new techniques or new data sources, particularly in a context marked by Big Data. The most secure anonymisation technique is still data aggregation, which transforms individual data into collective data. But these techniques make numerous subsequent processing operations impossible. It is therefore often legitimate to retain the individual nature of the data while controlling the risk that the persons concerned could be reidentified.

13. Opinion 05/2014 on Anonymisation Techniques.

14. According to the terms of Article 4(5) of the European Regulation, pseudonymisation is '*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*'.

Anonymisation (a reversible action) is desirable wherever possible and provides impersonal data. In all other cases, the individual data must be considered as pseudonymised (or indirectly identifiable) and present a more or less high risk of both reidentification and disclosure. It is the assessment of these two risks (reidentification and disclosure) in relation to the sensitivity of the data processed that must guide the implementation of appropriate security measures.

Is the distinction between pseudonymisation and anonymisation sufficient in relation to Big Data?

Data Mining, which involves the processing of large-scale data using powerful calculation techniques, uses either anonymised or pseudonymised data, and in very rare cases, if ever, directly identifiable data.

For certain studies, anonymised data can have limited value in relation to the intended objectives, and aggregation is insufficient to reflect the health phenomena that only more detailed data make it possible to reveal.

With regard to pseudonymisation, this technique can conflict with the current conditions for compliance with the rules protecting personal data as these are defined in France by the Law on Data Processing Data Files and Personal Freedoms and soon in Europe by the above-mentioned European Regulation.

How can we determine in advance the specific purpose for which we want to collect and process data when the factor that specifically characterises new research methods like Big Data is the processing of data without knowing the purpose in advance?

In addition to the need to comply with this principle, there is an obligation to complete the prior formalities imposed by the CNIL, which are, as we know, complex and lengthy. The CNIL also has the option of certifying the process associated with anonymisation of personal data, for the purpose in particular of reusing public information placed online[15].

15. Article 11 of Law No 78-17 of 6 January 1978, amended by Law No 2016-41 of 26

While the applicable legal texts do currently provide some options, these remain limited and do not enable consideration of actual requirements.

January 2016: '*It may certify or approve and publish reference guides or general methodologies for the purposes of certifying compliance with this law in relation to personal data anonymisation processes, in particular for the purposes of reuse of public information placed online under the conditions provided for in Title II of Book III of the Code of Administrative Procedure.*'

# ARTICLE VI
# RELEASING DATA: PATIENTS, USERS AT THE HEART OF THE 'DISRUPTION'

Thanks to the collection and analysis of Big Data, techniques for prevention, treatment, diagnosis and the monitoring of patients have been changing at an accelerated pace, since mobile health especially has been available to download on our smartphones.

The collection of these data, commonly referred to as Big Data, is most certainly at the point of transforming the methodology used by researchers and the approaches of the various institutional stakeholders involved in managing health risk and expenditure in France.

While the notion of Big Data is now the 'fun' translation of the existence of intensive and industrial-scale health data collection, it does not say anything about the procedures for accessing this information. For this, another English phrase is often used: Open Data. This concept, created out of a societal movement favouring maximum transparency and civic democracy, is a response to users' 'right to know'. Although that is not all it represents; Open Data in health is a philosophy based on free access to digital health data, irrespective of the author. It assumes the structured dissemination of those data, according to a method and an open licence guaranteeing free access and reuse by everyone, without technical, legal or financial restriction.

**Our law recently changed in order to modernise our health system and redefine, on that basis, a process providing access to health data held in the National Health Data System (SNDS),** made up of the following databases:

• the data produced as part of the Information Systems Medicalisation Programme (PMSI), which reflects the activities of both private and public healthcare institutions;

• the data from the National Interscheme Health Insurance Information System (SNIIRAM) produced by the entities responsible for

managing a basic health insurance plan;

• the National Register of Causes of Death;

• the data produced by Departmental Centres for Disabled People subject to the authority of the Independent Living Support Fund (CNSA);

• a representative sample of reimbursement data by beneficiary provided by top-up health insurance entities and defined jointly with their representatives.

The purpose of the SNDS is to provide data to contribute to:

• information about health and treatment options, medical and welfare management and associated quality;

• the definition, implementation and assessment of health and social welfare policies;

• an understanding of health expenditure, health insurance expenditure and medical and welfare expenditure;

• information from health, medical and welfare professionals, structures and institutions about their activities;

• health surveillance, monitoring and security;

• research, studies, evaluation and innovation in the areas of health, medicine and welfare.

The data from the SNDS that are made available to the public are processed into aggregated statistics or individual data created in such a way that the direct or indirect identification of the persons concerned is impossible. These data are made available at no cost. The reuse of these data cannot have the object or effect of identifying the persons concerned.

Access to the personal data held by the SNDS by research and analysis institutions operating for profit (and in particular entities that produce or market healthcare products, credit institutions, companies undertaking direct insurance or reinsurance activities and insurance intermediaries) may only be authorised to enable processing for the purposes of research, analysis or evaluation contributing to one of the six purposes listed above and corresponding to a public policy reason. In

this context, access is only authorised 'to the extent that these actions are rendered strictly necessary by the purposes of research, analysis or evaluation or by the missions entrusted to the body concerned', and only the data necessary for that processing may be used. Beyond that, a specific request for authorisation must be made to the CNIL.

The new National Institute of Health Data (INDS) receives requests for authorisation, which must then be submitted for an opinion from the expert committee for research, analysis and evaluation in the area of health. This committee will announce a decision as to the methodology chosen, the need for the use of personal data, the relevance of those data in relation to the intended purpose and, where applicable, the scientific quality of the project. At the initiative of the CNIL or at its own initiative, an opinion may be requested from the INDS in relation to the public interest nature of the research, analysis or evaluation justifying the request for processing. The CNIL will then assess the project in terms of the principles relating to protection of personal data and the value represented by the request.

The INDS will publish the authorisation from the CNIL, the declaration of interest and the results and method.

The three main associations representing users (UNAF, FNATH and CISS) have noted the need to include representatives of patients and users on the strategic and monitoring committees to be created.

This is therefore how the French law envisages the operational dimension of Big Data in health, taking the view that research projects implemented by profit-making entities must be subject to additional constraints in order to access the data held by the SNDS because of their commercial focus.

Press release from HDI, September 2016

## ACCESSING HEALTH DATA IN FRANCE IS A LONG, COMPLEX AND UNEQUAL PROCESS

**Paris, 26/09/16 - The Law of 26 January 2016 reforming our health system demonstrates a commitment to opening access to health data, but struggles to convince as to the objectives that France should be seeking in this regard. The Healthcare Data Institute would like to draw attention to the complexity of the mechanism introduced and has drawn up certain proposals to resolve this situation.**

The processing of health data is becoming a potential source of considerable value creation, in particular for scientific research and the acquisition of new knowledge, from the management of care to the definition of health policies By creating a new Title VI devoted to the provision of health data, the Law of 26 January 2016 reforming our healthcare system seems to be demonstrating serious ambitions.

'*It should be noted that, in the current version of the text and before publication of its implementing decrees, the mechanism introduced does not currently appear to be operational or able to allow the opening up of the data that it has heralded, creating an onerous and complex procedure, particularly for companies in the private sector*', explains Isabelle Hilali, Chair of the Healthcare Data Institute.

### Increased complexity and absence of visibility

The members of the Healthcare Data Institute have observed an increased level of complexity since the vote on the Law of 26 January 2016.

At least six steps must be overcome to obtain authorisation to implement processing for research, and no less than six authorities may intervene in the process (INDS, CNAMTS, CNIL, Expert Committee, research laboratory or agency, trusted third party, etc.).

'*The stakeholders, whether institutional or economic, need stability in order to develop their missions and their business. It is difficult to lead a industrial project or a major national project smoothly and with equanimity in a legal context that is changing in reverse of other texts and, above all, against the current tide of digital technologies*', notes Jeanne Bossi

Malafosse, barrister, Counsel for DLA PIPER and member of the Healthcare Data Institute.

### Inequalities in access to data

For the members of the Healthcare Data Institute, inequalities in access to data still exist in the new conditions introduced by Article 193 of the Law of 26 January 2016. These demonstrate a lack of trust, or indeed a mistrust, of private players, which are nonetheless at the very heart of innovation in the healthcare sector, as has always been the case.

The potential risk is reflected in a decrease in France's competitiveness in terms of analysis of health data. Certain studies are currently being conducted using foreign data, which are easier to access, and no infringements of rights and personal freedoms have been observed. There is also a risk that parallel databases could develop, thus undermining national ones, which have considerable scientific value.

Let us hope that the implementing texts for the law will introduce a governance of the National Health Data System that is open, at a time when the benefits come from the combined contribution of all private and public stakeholders.

**The above considerations have therefore led the members of the Healthcare Data Institute to request a review of these provisions.**

### Four proposals put forward by the Healthcare Data Institute

• **Align** the procedures between public and private stakeholders where a public health interest is being pursued and where appropriate guarantees are in place.

• **Apply** a process of control based on the principles that now drive the protection of personal data in all European Union Member States (*Accountability and Privacy by Design*): control *a priori* streamlined and defined on the basis of an analysis of risk, control *a posteriori* reinforced, much heavier penalties in the case of non-compliance. In this regard, the possibilities for penalties imposed by the CNIL must be reinforced.

• **Distinguish** according to the type of data, in order to ensure that the same constraints are not applied to data that have only a minimal risk of reidentification. In this regard, we must not confuse the intensity of the risk (the consequences of its occurrence) with the probability of the occurrence of that risk.

• **Ensure that the possibilities** recognised by the texts for the CNIL to simplify the procedures (single authorisations, reference methodologies, etc.) are actually effective.

> Read the detailed position of the Healthcare Data Institute.

## But will the SNDS really be the holy shrine of health data?

Thanks to the numerous smartphone sensors and the rampant spread of health and well-being applications, health data are being disseminated throughout a landscape that is much more vast and impressionistic than just the SNDS, whose administrative design we have just discussed.

The traditional proponents of Big Data, in particular the French National Health Insurance Fund for Salaried Workers (CNAMTS) and healthcare institutions, because of their coding, in fact no longer have a monopoly on health data. This more widespread access has less to do with Open Data than with the wide range of entities offering electronic services. Health data can take many different forms: medical and administrative data in an original approach; raw and meaningful data when they are prerequisites for apps that make it possible to monitor physiological vital signs; to geolocate users in order to position them in relation to a health institution; and to improve treatment compliance, when this does not involve proposing a slimming coach, a digital companion for better ageing, a connected night attendant, etc.

One of the most pressing questions in relation to the data collected in the context of the *quantified self* relates to the future of those data. What will become of these masses of information collected with the more or less informed consent of the users? Who will they actually profit? And how can recourse to the concept of Open Data be valuable for these new uses, which we have been describing as such for

several years now?

The proportion of data provided by personal sensors out of all information stored is expected to increase from 10% to nearly 90% over the course of the next decade.

The challenges for the holders and hosts of all these data will be measured on the basis of the profitability that these data represent for the promoters of enhanced health and well-being, but not exclusively. Because while the value of health data exceeds the value of banking data on the 'black market', this is because they have a particular appeal for scrupulous and benevolent major accounts.

After the cyberattacks sustained by medical analysis laboratories or health institutions, will the providers hosting the data required by health apps be the next to be held to ransom?

**Is the positive law robust enough to guarantee the security of the 'shattered' health data?** More generally, are we equipped to take advantage, collectively and individually, of Big Data while protecting individual rights?

The new European Regulation on the protection of personal data, published in the Official Journal of the European Union on 4 May 2016, which will enter into force in 2018, should make it possible for Europe to adapt to the new realities of the digital world, by recognising the following in particular:

• The obligation to provide clear, intelligible, easily accessible information for the persons concerned by data processing operations.

• The need for stated consent: users must be informed of how their data will be used and must in principle provide their consent for processing those data, or they must be able to object to processing. The responsibility for proving consent falls to the data controller. The realisation of this consent must not be ambiguous.

• The right to data portability making it possible for a person to recover the data that he or she has provided in a form that can easily be reused, and, where applicable, to then transfer it to a third party. This involves giving people back the control of their data, and offsetting some of the asymmetry between the data controller and the person concerned.

• The right, for associations active in the area of protecting personal rights and freedoms in relation to data protection, to introduce collective redress in respect of the protection of personal data.

• The right, for any person who has sustained material or non-material injury as a result of an infringement of this Regulation, to obtain redress for the loss sustained from the data controller or subcontractor.

This European Regulation is thus pushing forward the right of access by individuals towards a right to portability by virtue of which they now have a 'right of return' and a reappropriation of data relating to them. Big Data is therefore becoming valuable for individuals, who are becoming in part 'deciders' in relation to how their data are used. Using this precious resource, there are numerous services that could one day offer solutions with significant added value for individuals, who have rightly been repositioned at the centre of these flows.

**These new approaches, which are highly personalised, will reorientate research towards solutions that are suited for individuals, their lifestyles and their constraints, and will enable a sensitive and detailed definition of their micro-objectives and motivational methods.**

**More collectively, data from Big Data, Open Data and Self Data could prove essential for epidemiological research,** because they will enable professionals to better understand the health status of the population, to adjust the treatment administered to the patient by observing models developed on a much broader scale, or to draw new conclusions, for example in terms of the relationship between the course of an illness and the associated environmental factors.

In addition, better use of health data could generate improvements in productivity and cost reductions within the healthcare sector (in the United States, the forecasts are for 300 billion dollars, in value, per year).

There is still enormous uncertainty about how Big Data could change our understanding of our relationship with health and how we can rationally and efficiently drive our health policy. But there is reason to believe that we are standing on the brink of a fundamental revolution that some people are referring to as a 'disruption', supported by the digitisation of our service economy, by the prosperity of miniaturisation and by the substantial capitalisation of IT companies that bring in

their wake numerous innovative start-ups.

The rights of individuals must be seen as valuable resources in this new economy that could be dominated tomorrow by the science of algorithms. In any case, this is the basis on which Big Data will benefit us all, and primarily the individuals themselves.

## ABOUT THE
## HEALTHCARE DATA INSTITUTE

Created in 2014, the Healthcare Data Institute is the first international Think Tank dedicated to Big Data in the area of health, and acts as a catalyst for ideas and projects focused around Big Data within the health ecosystem.

The Board of Directors of the Healthcare Data Institute includes representatives of Aviesan, Axa, Caisse des Dépôts et Consignation, CEA, CRI, Groupe Elsan, McKinsey & Company, Orange Healthcare, Quintiles-IMS and Sanofi.

HEALTHCARE
**DATA** INSTITUTE


**CONTACT**
Pierre-Yves Arnoux
office@healthcaredatainstitute.com
+33 (0)6 07 13 77 13

**21, rue Jasmin**
**75016 PARIS - FRANCE**

@HCDATAINSTITUTE
healthcaredatainstitute.com