Cohort Innovation Day – October 3rd, 2018

Key takeaways and feedbacks from this one-day event

France is rich in many public and private healthcare data sets, some of which are already in use for various applications. Moreover, France is developing a national strategy on Artificial Intelligence (AI) presented by the French President Emmanuel Macron in March 2018, following the report written by Cédric Villani. This national strategy in AI led to major ongoing initiatives in the health sector, especially the Health Data Hub prefiguration mission and the Strategic Committee for Industry Sectors (Comité Stratégique de Filières, CSF).

In this context, ARIIS (Alliance for Research and Innovation for Health Industries) and Aviesan (Alliance for Life Sciences and Health) organized on October 3rd, 2018 the third edition of the Cohort Innovation Day, an event focusing on epidemiological data and cohorts. The 2018 edition puts special focus on the role of AI in leveraging the power of French cohorts and transforming medical research. The aim of the Cohort Innovation Day was thus to foster interaction between cohorts as high-quality databases and private actors developing AI. It was also an opportunity to grasp what role the Health Data Hub might play as a coherent and holistic health data warehouse.

This paper gives an overview of the main ideas and opinions presented during the debates and round tables of the Cohort Innovation Day.

1. THE NECESSITY OF STRUCTURED AND MASSIVE AMOUNTS OF DATA

"The era of revolutionary discoveries is over, we are now facing narrower research domains, exploring massive and various data flows" *by Gilles BABINET, France's Digital Champion for the European Commission*

The medical and life sciences world is in an international context where revolutionary discoveries are becoming scarce. **Scientific research requires now massive amounts of data and large observatories** for exploring specific medical fields or risks with lower magnitudes. Thus, setting up an efficient healthcare AI strategy requires first and foremost the assembly of structured and massive amounts of data.

In this context, France is rich in different cohorts, mostly patient-based (pathology related) but less so in large population. Nevertheless, French cohorts are a goldmine for high-quality health data and can be considered **international references in terms of quality-by-design approaches** allowing to easily set-up and train machine learning algorithms on nearly-real-world-data. Leveraging the richness of French cohorts promises to decidedly propel AI further and help clinical research.

➢ See appendix 1 (p.5): The **Constances cohort** ant the **integrative iBox system**

Despite this high-quality amount of data and the potential generated by cohorts, **security framework** and **data access** are still challenges that need to be addressed, with no standardized answer so far. France is indeed suffering from *distrust* regarding data sharing that tends to put the country offside compared to other countries offering more possibilities to combine data (e.g. the UK Biobank). This complex access to data is also detrimental to the development of the French panorama of start-ups: the successful start-ups developing AI related technologies for diagnostics and prevention products and relying on massive sets of data often seek their data abroad. The **players's willingness to share their data** is thus crucial to truly leverage the power of AI.

In this regard, the **Health Data Hub** will play a key role by creating a "*single desk*" for all the available data in France. The first use cases defined through the *Strategic Commitee for Industry Sectors* (CSF) and intended to use this Hub will also be a real opportunity to define a durable business model for data sharing and simplify collaboration with French cohorts.

2. AI IS A POWERFUL TOOL FOR BIOMARKER DISCOVERY

"Many projects are built around AI and raise high hopes for medical applications" *by Xosé FERNANDEZ, Chief Data Officer, Curie Institute*

Hypothesis-driven research has been extremely fruitful for generating discoveries in science including medicine. However, this linear process – very useful for choosing between several mechanisms – has limitations, especially when dealing with complex systems such as biology, and may be biased. The increase in computing power makes possible to handle large volumes of data leading to the rise of **data intensive science** (data-driven approaches) that may help overcome hypothesis-driven approaches' limitation through unbiased (non-supervised) methods and therefore open avenues for novel discoveries including disease management and drug treatment. With AI and big data analytics, it is now more than ever possible to link symptoms or conditions to diseases. AI is therefore creating opportunities to **transform and improve the healthcare system**, in a very natural way and in many aspects:

- **Improve treatment discovery** and application,
- **Enhance patients' response** to treatment,
- Make **personalized medicine** with combined therapies,
- Fill the void created by **lack of physicians** in remote areas,
- Avoid **unnecessary hospitalizations**,
- **Transform the role of physician**, not by taking away the final decision made by physicians but by using data and AI to improve such decision process (efficiency, reliability) to finally improve patients' health.

Many players (private and public) are currently leading different projects applying AI to large data sets for new biomarker discovery.

➢ See appendix 2 (p.5-6): Presentation of the **PACIFIC consortium** and projects led by **Philips,** the **Institute for Brain and Spinal Cord** (ICM) and **Owkin.**

However, AI and more generally **data-driven methods** alone are not sufficient *per se* and need to be **combined with the hypothesis-driven approach** to harness their full potential. In addition, AI is still in its infancy and some challenges remain to be overcome prior to its full exploitation. Studies must and will be led on techniques allowing for multiple data-type analysis in a single model (images, texts, etc.).

3. AI CAN LEVERAGE THE POWER OF REAL-WORLD DATA TRANSFORMING IT INTO TOMORROW'S GOLD STANDARD

"Drugs always react differently in the real world, using real life data is a logical and necessary step to optimize drug development" *by Samir MEDJEBAR, Business Development Director, BioSerenity*

"A drug is more likely to be efficient in the real world when there is genetic evidence" *by Philippe SANSEAU, Head Computational Biology and Stats, GSK*

Today, **connected devices and products generate massive amounts of data** creating new tools for epidemiology and leading to potential new discoveries, identification of biomarkers, behavioral or social impacts, etc. These kinds of discoveries and behavioral patterns are creating a **new disruptive paradigm in population epidemiology**.

The pharmaceutical industry has today truly understood the **potential of real-world data analytics**, a field of study where we can see several disciplines converging. For instance, studies have shown that using genomic data leads to improved results in drug research. Since 2016, Sanofi has been working on introducing big data and trans analytic technologies in drug development. The current economic situation leads to believe that real-world analysis will first **support traditional R&D methods** but will most likely replace them in the long run. Efficiency and safety measures are the main challenges to overcome. Even regulatory agencies such as the FDA and the EMA have already or will soon publish guidance support on how to apply real-world data in health products R&D.

➢ See appendix 3 (p.6): Project led by **BioSerenity**

The **success of real-world data analytics** will thus depend on multiple factors, including:

- actors' willingness to share their data: holders of large data bases will most likely not give away their core value without any incentives;
- coherent and shared data structure and access;
- leveraging public-private partnerships.

4. INNOVATION : A DIFFICULT YET NECESSARY DRIVER IN HEALTHCARE

"The reason we are developing these innovations comes from a collision between social values and financial reality for maintaining health systems around the world" *by Ran BALICER, Founding Director, Clalit Research Institute (Israel)*

"Health systems will not be replaced by algorithms, but health systems that do not use algorithms will be replaced by those who do" *by Ran BALICER, Founding Director, Clalit Research Institute (Israel)*

Innovation in healthcare seems difficult because **healthcare is an industry that cannot sustain mistakes**, yet the current financial situation is forcing actors to evolve. Data analytics and AI gives us the power to **move from a reactive to a proactive state**. Yet, analyzing data can be challenging since data is based on the average population, and we all have our own specificities, "none of us is an average".

In France, well before the 21st century AI wave, the Ministry of Health launched in 2008 the premises of a global data collection and storing solution that was the DMP (Personalized Medical Record). Back then, the DMP had a planned road map that some may consider more ambitious than the current one: it was to become an epidemiologic tool as well as a personal health data record. However, in 2018, France seems to be slowly losing its innovative position in

the healthcare sector with many other countries far ahead in implementing personalized health records. One of the main reasons seem to be tied to the numerous regulatory frameworks and the non-unified data health governance that have been **slowing down French innovation**. For instance, the CNIL (National Commission for Information Technology and Civil Liberties) seems to have been over-protective with personal data management:

- In France, an individual's social security number cannot be used as a unified base code to cross reference multiple databases and cohorts;
- When building the SNDS (National Health Data System), the design of the database was focused on avoiding security risks rather than trying to identify ways to generate health benefits. The SNDS is therefore maladjusted for machine learning purposes;
- Healthcare data governance is complex and shared between different bodies: ministry of health, HAS, SNIIRAM, CNIL, etc.

These realities have given other countries strategic advantages even though the French health ecosystem is rich in medical databases and cohorts and holds high promise in developing disruptive AI technologies. For instance, in Israel, the Clalit healthcare system (over 1 400 clinics) is completely interoperable, and all data are recorded into a single system. The Israeli system holds decades of ID tagged, geo-coded data. Analysis of the data allows the system to predict, with high likelihood, patients who might suffer from kidney failure five years prior. The solution is adjusted to the local population and answers a real medical need with already impressive results.

In the current situation, **avoiding risk-taking might become a risk itself**. Private and public actors need to find ways to collaborate: cross-linking data is a must and will create a unique opportunity in disease understanding and patient management. To succeed, actors in France need to **address the issue of the economic model of data sharing** very fast. In a world that has been historically managed by clinicians, if innovative AI solutions hope to be used, they also need to address clinicians' specific needs: understanding the outcomes and integrating into their workflows.

The promises of AI are high, but the technology is still maturing. Data needs to be explored in a smart way through the creation of accessible and structured data sets. Succeeding in the development of a true holistic health data strategy in France will forever **change how patients and disease are treated**.

5. CONCLUSION

The Cohort Innovation Day is a unique opportunity to:

1. **gather all stakeholders**: tomorrow's challenges that promises to transform patient and population care management will only be addressed through a **holistic and shared health data strategy**. Thus, the Cohort Innovation Day gathers French and international industry leaders, top talented scientists, CEOs from start-ups and key opinion leaders. Debates contribute to promote the willingness to share data and to build a successful regulatory framework including all players/regulatory agency to encourage use. This 2018 edition has gathered 250 participants including 48% of industrial and 28% of public researchers.

2. **spotlight France competitiveness and attractiveness in research and innovation in life and health sciences**. France has strong research capabilities in AI and a pool of talents trained in highly regarded engineering schools, so that many large international groups have set up AI research centers in France (e.g. Philips, Google Deepmind, Samsung, Facebook). However, several countries including US, China and Israel created a favorable environment and are ahead. Others EU countries are also moving quickly. A change in mindset and a supportive environment are mandatory to leverage these capabilities.

3. **Promote public/private partnerships and develop the best conditions to enhance innovation**, especially in data and complex algorithm sharing. 84 B2B meetings between 11 industrial delegations and 16 cohorts have been organized during the Cohort Innovation Day. 10 promising start-ups in IA and health data took also part in this event and presented their technologies during a networking session.

To conclude, the ongoing initiatives including **the Health Data Hub** and the **Strategic Committee for Industry Sectors** are the first stepping stones in building this *health-data ecosystem* by creating a "single desk" under an unified governance and defining a business model for data sharing.

6. APPENDICES

APPENDIX 1:

- **The Constances cohort**, managed by the Population-based Epidemiologic Cohorts Unit - UMS 011 INSERM-UVSQ, is the largest cohort in France and is considered as a mega cohort (i.e. a cohort with over 100 000 participants), focusing on multiple diseases and general health state. During follow-up cohort members complete each year a personal questionnaire, and every four years, Constances participants go through a full medical check-up (clinical examination, blood tests, etc.), and their data are regularly extracted from the SNIIRAM-SNDS and CNAV databases. Participants residential addresses are geocoded so that environmental data can be cross linked to their personal data. Every data collecting process is precisely monitored (duration of fasting before a blood test, etc.) and the quality of newly integrated data is controlled. Constances currently holds about 80 000 cognitive functions reports, 170 000 ECGs and has more than 30 publications. The cohort will soon have 200 000 participants. Today, Constances is working on data-linkage projects with other cohorts and French data sets. The team is also preparing datasets from the SNDS and other public databases for AI use.

- **The integrative iBox system,** managed by Alexandre Loupy, Director of the Paris Transplant Group at Necker Hospital and Professor of Nephrology and epidemiology, is a kidney transplant survival prediction model based on an international 15-year-old cohort. It follows many kidney transplantees from their transplantation date until a potential graft removal. This data-based approach transforms cohorts into tools capable of identifying causality and new hypothesis in kidney transplant. Concrete results are already published and have led to improved prognostic assessments, personalized clinical trials, and better therapeutic efficacy. Algorithms can determine the nearest profile within the iBox cohort for a given patient. This reference profile from the cohort calculates rejection risks of an allograft transplant, predicts patient's recovery trajectory, and evaluates long term risks. Currently awaiting FDA approval, the concept could be applied to other medical domains.

APPENDIX 2:

- **The private-public PACIFIC consortium** (that associates Sanofi, Servier, APHP, as well as INSERM and several SMEs) is being set up to transform care delivered to the 50% of patients suffering from a kind of heart failure with no validated/specific treatment at this time. Based on deep phenotyping including clinical data, high-end biomarkers and imaging methods, as well as data generated from continuous monitoring equipments, AI (unsupervised machine learning) will help in clusterizing population. These clusters will then be used to develop specific drugs or reallocate existing drugs to new purposes to serve this population. This study will recruit 500 patients over a 5-year span and will allow pharmaceutical companies including Servier and other healthcare players to adopt data-driven strategies and provide better patient care.

- **Philips** recently announced the creation of an AI research center in France. The center will host an incubator and develop decisional tools accessible through an online platform for hospitals and physicians looking for diagnostic support, tools to evaluate treatment responses or monitor drug adherence. The platform has been selected to be used in genomic data analysis in the national genomic program *France Médecine Génomique 2025*.

- **The Institute for Brain and Spinal Cord (ICM)** is leading the development of a data warehouse for neuroscience research data. Based on the neurological assessments and medical images of several hundreds of patients, ICM with INRIA are working to create personalized artificial brain models that can early predict Alzheimer onset.

- **OWKIN** is an AI startup that builds machine learning tools to augment medical and biology research. The proprietary platform OWKIN Socrates integrates biomedical images, genomics and clinical data to discover biomarkers associated with treatment outcomes. OWKIN builds models that are as interpretable as possible, for a deeper understanding of the initial data. Through transfer learning, the start-ups pushs the intelligence of pre-trained algorithms into new models which will have a better predictive power from cohorts as small as 100 patients.

APPENDIX 3:

- **BioSerenity** is developing connected wearable medical devices that can produce high quality contextualized clinical data. The BioSerenity devices can tackle the issue of unstructured real-world data using reliable digital biomarkers. In a prospective point of view, a global use of reliable connected sensors could lead to real-time dynamic patient management as well as dynamic clinical research.

7. AUTHORS

This document has been written with the contribution of:

- **Michèle ARNOE,** Head Growth and Innovation, IQVIA
- **Gilles BABINET,** France's Digital Champion for the European Commission
- **Ran BALICER,** Founding Director, Clalit Research Institute (Israel)
- **Stanley DURRLEMAN,** Co-Head, INRIA/ICM Brain and Spine Institute, ARAMIS Lab
- **Xosé FERNANDEZ,** Chief Data Officer, Curie Institute
- **Marc FROUIN,** Chief Operating Officer, BioSerenity
- **Ramon HERNANDEZ,** Head of Advanced Analytics in R&D, Sanofi
- **Alexandre LOUPY,** Nephrologist, Necker Hospital, Paris Descartes University and Assistance Publique-Hôpitaux de Paris Inserm U970
- **Rosalie MAURISSE,** Head of the health sector, Expertise Direction, BpiFrance
- **Samir MEDJEBAR,** Business Development Director, BioSerenity
- **Meriem SEFTA,** Partnership and Translational Research Manager, Owkin
- **Philippe SANSEAU,** Head Computational Biology and Stats, GSK (UK)
- **Benoit TYL,** Director of Cardiovascular Translational Research, Servier
- **Nicolas VILLAIN,** Head of Research in AI for Medical Imaging, Philips