Initiative “économie de la santé” / “Health economics” initiative

Big pharma, big tech and the ecosystem of health data production:

Is the emergence of big data and AI about to transform the structure of the health industry?

Context

The research project aims to develop an original approach on the transformations of the health sector related to the development of artificial intelligence and big data. While the main academic field will be economics, interdisciplinary approaches are more than welcome, in particular if they involve institutions belonging to the *Sorbonne Université alliance* (faculté de médecine, iPLESP, INSEAD, etc.). If relevant for the applicant, a joint supervision with another researcher in France or abroad might be considered.

The PhD will be supervised by David Flacher at the COSTECH lab, *Université de technologie de Compiègne*, member of the *Sorbonne Université alliance*. His field of research includes innovation economics and the economics of the digital economy. The PhD will be jointly supervised with Nathalie Coutinet (associated member at COSTECH, UTC), as a specialist of health economics (author, among many other scientific articles and book chapters of “Economie du médicament”, 2018, with Philippe Abecassis, Repères, La découverte)

Research proposal

The emergence of big data and artificial intelligence (AI) are potential sources of major transformations of the health industry. On one side, big pharma companies including major players, mostly in Europe and the US, are facing the need to incorporate growing amounts of health data in their R&D and product development strategies. On the other side, big tech companies – the GAFAM (Google, Amazon, Facebook, Apple, Microsoft) and the BATHX (Baidu, Alibaba, Tencent, Huawei, Xiaomi) – are concentrating an increasing amount of data from almost all industrial fields. Thanks to this concentration of data, they could become the major player in AI.

All along the global value chain (GVC), the economic activity is increasingly influenced by data, which may offer an increasing power of big tech companies over various industries. The extent to which big pharma companies are threatened by big tech firms is thus a major issue: from a theoretical perspective, it raises the questions of the new forms of interdependence between firms and sectors, the new nature of the firms and the new industrial frontiers in this new context. It is also a major topical issue, as far as the leading role taken by the big tech companies has led to a mounting pressure to mobilise the antitrust law in both Europe and the US. Tech giants span all the way from infrastructure to applications, also leading much of the relevant scientific advantage (Jacobides, Brusoni, and Candelon 2021). These mostly vertically integrated companies are driven to stimulate the production of AI (and trumpet its benefits), as they stand to benefit from the improvement in their products. Exactly for this reason, a large portion of the required investment has been sponsored directly by them (Jacobides, Brusoni, and Candelon 2021). This is a critical point, particularly considering concerns about basic research's dwindling significance in corporate R&D. We face a unique dilemma in AI, as research is dominated by a handful of firms. The development of big data and AI has also led to the emergence of a large number of start-ups and bigger firms composing the “health tech” sector.

Within this context, a typology of the different players appears a key stake to understand the main ongoing transformations in the health industry and in particular to analyse who is producing the data

and tools, who are funding these developments, how they serve or hamper the strategies of big tech and big pharma companies.

In particular, the breakthrough technologies, which are part of a new innovative concept called the Internet-of-Medical-Things (IoMT) (Islam et al. 2015), have the potential to transform the healthcare industry. In the smart healthcare environment, IoMT technologies contribute to the collecting of medical data for intelligent data analytics enabled by artificial intelligence (AI) (Nguyen et al. 2021) and assist in the development of a variety of innovative smart healthcare applications, ranging from cancer biomarker detection to patient screening and genetic prediction via imaging. As artificial intelligence depends on constant flows of data to recognize patterns, and then make predictions, it needs access to large and diverse datasets to train, improve their accuracy, and eliminate bias.

As for now, the centralisation of data has proven to be key to develop the most advanced AI algorithm. This can be seen as a central feature of the power gathered by the big tech companies, with the consequence of threatening many other economic sectors and leading to very concentrated oligopolies. In addition, the pooling of data especially creates a challenge for the health industry, as it must be balanced with critical concerns regarding patient privacy and data protection. This is especially true in e-healthcare, where health-related data is very sensitive and private, subject to special rules (Cheng and Hung 2006). Nevertheless, because in the future health data could potentially be disseminated throughout a large-scale IoMT network, such a centralized AI architecture may no longer be suitable for healthcare systems (Nguyen et al. 2022). Hence, moving toward distributed AI techniques for scalable and privacy-preserving intelligent healthcare applications at the network edge is crucial.

It appears that alternative approaches to AI and big data could be considered. Innovative unconventional learning paradigms constitute examples of how data dependence could be restructured. A successful implementation of federated and transfer learning, could hold significant potential for training machine learning models across multiple medical institutions and enabling precision medicine at a large-scale; helping match the right treatment to the right patient at the right time. Algorithms may employ what they've learned from a data source not only to generate better predictions in respect to that data's underlying problem, but also to transfer that knowledge to other cases. Extracting knowledge from one or more application scenarios to help boost the learning performance in others is called transfer learning (TL). Compared to traditional machine learning, which requires large amounts of well-defined training data as an input, TL can help promote AI in less-developed application areas (Yang et al. 2020). As mentioned by Yang et al. (2020), examples of the use of TL, include (but are not limited to) building a recommendation system in a new domain (with insufficient training data) with the help of a related well-developed but different domain (e.g. books and movies). In general, it can potentially broaden the scope of Internet of Medical Things (IoMT) businesses since AI models trained with data from one region could be transferred to unexplored other regions, within the same disease group (e.g. illnesses caused by similar viruses) or from one hospital to others.

Another learning paradigm that addresses the issue of centralizing data by enabling collaborative learning is federated learning (FL) (Rieke et al. 2020). In a FL setting, algorithms learn while processing data that do not need to be hosted together at the same location. Federated learning models are trained in a distributed manner without having to share the underlying data, allowing algorithms to be distributed to several data centres. In these locations, they receive local training, and only the information gathered by the algorithm is sent back, while better predictions are then handed over to the local datasets to be re-trained and enhanced. FL is a distributed AI technique that could potentially allow high-quality AI models to be trained by averaging local updates from several health data clients, such as IoMT devices, without requiring direct access to the local data (Nguyen et al. 2021). Federated and transfer learning have emerged as a promising decentralized alternative for implementing cost-effective smart healthcare systems with increased privacy protection (Warnat-Herresthal et al. 2021; Sheller et al. 2020; Kaissis et al. 2021). They move attention from data centralization towards data

access, proving that the former is not required to successfully train AI models. They can serve as the pioneers of tomorrow's privacy-enhancing, scalable smart healthcare networks and applications, that restrict the disclosure of sensitive user information and preferences, mitigating the risks of leakage.

The application of alternative learning techniques and decentralised machine learning in healthcare is still at its infancy and may play a crucial role in the development of large-scale, collaborative healthcare systems, allowing for a shift from centralized health data analytics to distributed healthcare operations while maintaining privacy. However, as mentioned by Jacobides et al. (2021), entire new subfields, such as federated and transfer learning, have essentially been created and thus dominated by companies, such as Google. Therefore, Big Tech are the incumbents, possessing -and keeping secret - the most ground-breaking AI models (Rikap and Lundvall 2021).

In the analysis of the impact of the emergence of big data and AI on the structure of the health industry, the PhD project will consider the role of the different players (big pharma, tech giants, start-ups of the heath-tech sector, etc.) of the health industry ecosystem on the R&D and product development strategies. It shall also analyse the relationships between these players and their respective impact on the main ongoing transformations. The research may also address the learning paradigms, the technological opportunities and public policy in relation to these major transformations.

Methodology and timeline

The PhD candidate will be asked to develop its own methodology. However, a few directions can be suggested already. They include of course an in-depth literature review of the scientific articles in the fields and an analysis of the grey literature (firms' annual reports, reports issued by national or international institutions, websites…) in order to build a typology of the economic conditions and strategies at stake for both the big pharma and big tech companies. It shall also be the basis of a theoretical discussion on the evolution of the markets. The analysis of the innovation strategies might also include the analysis of patents claimed by those companies and/or their scientific publications. Different tools could be used, such as network analysis, factorial analysis, econometrics, depending on the considered questions. This work may also be complemented by interviews of the main players and users in the field, for a more qualitative approach.

The *Derwent innovation* database, available at UTC, allows to gather clean databases of patents, while *Web of Science* allows to retrieve data on scientific publications. Accountability data for the considered firms can also be retrieved from databases such as *Datastream*, while more macroeconomic data can be obtained from OECD or ILO databases.

Indicative timeline

	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semester 6
Literature review + 1st conference paper						
Training on the tools the PhD applicant may not know						
Development of the theoretical framework						
Data collections (patents, publications, figures…)						
Data analysis						
Article 1 (journal + conf)						
Article 2 (journal + conf)						
Article 3 (journal + conf)						
Redaction and defence						

References

Cheng, Vivying SY, and Patrick CK Hung. 2006. 'Health Insurance Portability and Accountability Act (HIPPA) Compliant Access Control Model for Web Services'. *International Journal of Healthcare Information Systems and Informatics (IJHISI)* 1 (1): 22–39.

Islam, SM Riazul, Daehan Kwak, MD Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. 2015. 'The Internet of Things for Health Care: A Comprehensive Survey'. *IEEE Access* 3: 678–708.

Jacobides, Michael G., Stefano Brusoni, and Francois Candelon. 2021. 'The Evolutionary Dynamics of the Artificial Intelligence Ecosystem'. *Strategy Science* 6 (4): 412–35.

Kaissis, Georgios, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jungmann, and Marc-Matthias Steinborn. 2021. 'End-to-End Privacy Preserving Deep Learning on Multi-Institutional Medical Imaging'. *Nature Machine Intelligence* 3 (6): 473–84.

Nguyen, Dinh C., Ming Ding, Pubudu N. Pathirana, Aruna Seneviratne, Jun Li, Dusit Niyato, and H. Vincent Poor. 2021. 'Federated Learning for Industrial Internet of Things in Future Industries'. *IEEE Wireless Communications*.

Nguyen, Dinh C., Quoc-Viet Pham, Pubudu N. Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. 2022. 'Federated Learning for Smart Healthcare: A Survey'. *ACM Computing Surveys (CSUR)* 55 (3): 1–37.

Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, and Klaus Maier-Hein. 2020. 'The Future of Digital Health with Federated Learning'. *NPJ Digital Medicine* 3 (1): 1–7.

Rikap, Cecilia, and Bengt-Åke Lundvall. 2021. 'The Digital Innovation Race'. *Springer Books*.

Sheller, Micah J., Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, and Rivka R. Colen. 2020. 'Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data'. *Scientific Reports* 10 (1): 1–12.

Warnat-Herresthal, Stefanie, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, and N. Ahmad Aziz. 2021. 'Swarm Learning for Decentralized and Confidential Clinical Machine Learning'. *Nature* 594 (7862): 265–70.

Yang, Qiang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. 2020. *Transfer Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139061773.