Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220390

AP-HP Health Data Space (AHDS) to the Test of the Covid-19 Pandemic

Christel DANIEL[a,b,1], Nicolas PARIS[a], Olivier PIERRE[a], Nicolas GRIFFON[a,b], Stéphane BREANT[a], Nina ORLOVA[a], Patricia SERRE[a], Damien LEPROVOST[a,b], Stéphane DENGLOS[a], Alexandre MOUCHET[a], Julien DUBIEL[a], Rafael GOZLAN[a], Gilles CHATELLIER[a], Romain BEY[a], Marie FRANK[c], Claire HASSEN-KHODJA[d], Marie-France MAMZER[e], Martin HILKA[a] and AP-HP Covid CDW Initiative

[a] *Innovation and Data, IT Department, AP-HP, Paris, France*
[b] *LIMICS, F-75006, Paris, France; Sorbonne Universités, Paris, France*
[c] *Department of Medical Information (DIM), Hôpitaux Universitaires Paris-Saclay, AP-HP Paris, France*
[d] *Direction de la Recherche Clinique et de l'Innovation, Paris, France*
[e] *Université de Paris, AP-HP, Paris, France*

Abstract. Sharing observational and interventional health data within a common data space enables university hospitals to leverage such data for biomedical discovery and moving towards a learning health system. Objective: To describe the AP-HP Health Data Space (AHDS) and the IT services supporting piloting, research, innovation and patient care. Methods: Built on three pillars – governance and ethics, technology and valorization – the AHDS and its major component, the Clinical Data Warehouse (CDW) have been developed since 2015. Results: The AP-HP CDW has been made available at scale to AP-HP both healthcare professionals and public or private partners in January 2017. Supported by an institutional secured and high-performance cloud and an ecosystem of tools, mostly open source, the AHDS integrates a large amount of massive healthcare data collected during care and research activities. As of December 2021, the AHDS operates the electronic data capture for almost +840 clinical trials sponsored by AP-HP, the CDW is enabling the processing of health data from more than 11 million patients and generated +200 secondary data marts from IRB authorized research projects. During the Covid-19 pandemic, AHDS has had to evolve quickly to support administrative professionals and caregivers heavily involved in the reorganization of both patient care and biomedical research. Conclusion: The AP-HP Data Space is a key facilitator for data-driven evidence generation and making the health system more efficient and personalized.

Keywords. Data space, Clinical Data Warehouse, Electronic Health Record, Real-world data, Covid-19 pandemics.

1. Introduction

The analysis of "Real world" data (RWD) has generated important medical discoveries, especially in areas where traditional clinical trials would be unethical or infeasible [1] and represents an opportunity to optimize clinical trials [2,3]. In the era of big data, the

[1] Corresponding Author, Christel DANIEL, Innovation and Data, IT Department, AP-HP, Paris, France; E-mail:christel.daniel@aphp.fr

rise of data science and in particular of new Artificial Intelligence (AI) technologies e.g Machine Learning (ML), RWD is increasingly used to develop innovative digital technologies and new services supporting various activities of health professionals, in research, care and training [4]. RWD and AI should enable healthcare professionals to objectively and continuously measure what is being done, to compare the results of care over time and between health establishments, to question the gaps, to simulate and evaluate the impacts of certain strategic decisions on organizations or the quality of care.

In this context, the objective of this paper is to describe the AP-HP Health Data Space (AHDS) and the IT services that enable the reuse of observational and interventional health data to support clinical and translational research and to move towards a learning health system [5,6].

2. Material and Methods

Assistance Publique - Hôpitaux de Paris (AP-HP) is the largest university hospital in Europe with 39 hospitals (20,500 beds), mainly located in Greater Paris region, conducting more than 3,500 ongoing research projects all sponsors combined. Built on three pillars – **governance and ethics, technology and value proposition** – the AHDS has been developed since 2015 and its major component, the Clinical Data Warehouse (CDW) (https://eds.aphp.fr), made available at scale to health professionals and researchers in January 2017.

Given the importance of the ethical and societal issues raised by the rapid increase of digitalization and data-driven approaches, a **specific governance** has been set up at the strategic and operational levels in order to define the data access and use policy and ensure its implementation guaranteeing patients' rights and privacy.

Supported by an institutional **secured and high-performance cloud and an ecosystem** of tools, mostly open source, the AHDS integrates a large amount of massive healthcare data collected during care and research activities. Initially focused on a core set of data collected through the common Electronic Health Record (EHR)(ORBIS), the Extract-Load-Transform (ETL) process of the AP-HP CDW is now extended to various additional sources including legacy AP-HP applications as well as external sources in order to get a 360° view of patients' health status. Big data such as medical images – and soon genetic sequencing data – are not exhaustively duplicated within the AHDS but copied on demand if needed for any research project.

Beyond a **costly investment** in hardware, software, the implementation of the infrastructure and services supporting the secure and efficient processing of high throughput health data the ADHS also requires the collaboration of experts in sufficient numbers in various domains: clinicians, epidemiologists, IT engineers, biostatisticians and data scientists. Different funding approaches are currently under consideration to sustain the AHDS.

3. Results

The **governance structure of the AHDS** is based on a Steering Board and an Institutional Review Board bringing (IRB00011591) whose members are healthcare professionals, researchers, directors, patient representatives and persons qualified in ethics and that assesses all research projects requiring the use of AP-HP CDW data.

Patients cared at AP-HP are informed individually, via an explicit mention on all medical notes and patient booklets, of both existence and purposes of the CDW as well as the procedures for opting out. The rules of governance as well as the list of the studies carried out or in progress are available on the AHDS web site (https://eds.aphp.fr). In order to guarantee patient privacy, de-identification processes, initially implemented on structured data and extended to medical images and textual documents, are assessed by both the AP- HP Chief Information Security Officer (CISO) and Data Protection Officer (DPO). The **technical pillar** relies on the capacity of the Big Data platform the AHDS offers a set of services supporting data access and use. The AHDS operates large scale clinical trials and cohorts and an increasing number of real-world studies. Most of the interventional data of clinical trials and cohorts sponsored by AP-HP are captured using the institutional electronic data capture (EDC) system (CleanWeb). The recent deployment of the open source EDC system RedCap resulted in new research projects. Observational data are stored in the AP-HP CDW, enabling, as of December 2021, the processing of health data from more than 11 million patients (Table 1). The integration of various data from heterogeneous systems developed in silos, relies on a standardization process based on standard models - such as the OMOP model of the OHDSI project and HL7 FHIR - and reference terminologies – such as ICD10, LOINC, ATC, etc. that contributes to the FAIRification of the data (i.e to make the data Findable, Accessible, Interoperable and Reusable).

Table 1. Content of the AP-HP CDW (as of December 31, 2021). Focus on the research data mart in OMOP format.

Data category	Year of first integration (or expected year)
Demographic data	2017
Vital status (from a national data base)	2021
Clinical pathway and care sites	2019
DRG codes	2017 (ORBIS hospitals), 2022 (all)
Clinical documents	2017 (ORBIS hospitals), 2022 (all)
Vital signs	2017
Consultation and operative schedule	2021
Lab results	2017 (ORBIS hospitals), 2022 (all)
Anatomic pathology precedures	2020
Diagnostic codes, digital slides	2022
Radiological procedures	2021
Medial imagine (DICOM)	2017
Drug prescription/administration	2019 (ORBIS hospitals), 2022 (all)
Care plan	2017
Problems, family and personal history	2022
Covid (coronaOMOP)	2020
Cancer (drug/chemotherapy, prescription,multidisciplinary meeting reports)	2022
Patient reported outcomes (PROMS)	2022
Obstetrics, Infectious diseases, Nephrology, Pediatrics, Rare disease, Emergency care, etc.	
Social history	2023
Exposome : Air pollution, socio-economic data	2022

Between January 1, 2017, and December 31, 2021, 216 research projects have been approved by the IRB. Investigators used the AP-HP cohort builders (i2b2 since January 2017 and now the cohort360 open-source application developed in 2021) to execute direct queries on the CDW and generate secondary data marts. The detailed list of studies is available on the AP-HP CDW web site; their number and nature are described in Table 2. Research projects have significantly increased since 2017 in conjunction with the

involvement of additional data scientist recruited with Covid-dedicated funds or disease-specific donation (e.g., in the domain of cancer). Epidemiological studies performed through the CDW were designed with various purposes: for fine-grained analysis of specific phenotypes, to confirm known disease risk factors or consisted in diagnostic, prognostic or medico-economic studies (table 2). Almost half of the projects were performed using AI/ML approaches particularly in the medical imaging domain. Some of these projects have a clear objective of digital innovation development/assessment (n=61).

Table 2. Number and type of research and innovation projects based on the AP-HP CDW between January 1, 2017 and December 31, 2021

Year	Number of projects	Descriptive	Etiology	Diagnostic	Prognostic	Medico-economical	Digital innovation	Publications
2017	12	3	3	1	4		4	4
2018	26	11	2	5	7	1	13	4
2019	34	17		6	12	1	12	1
2020	95	46	4	11	35	1	18	11
2021	49	29	4	10	10		14	Nd

The also AHDS supports real-world studies implemented at national (French Health Data Hub (HDH)) or international scale (EHDEN or EHR2EDC consortium [7]. Thanks to the Pilote software application, based on the Cognos (IBM) solution, the AP-HP CDW data is also used for strategic and medico-economic management.

Regarding the **valorization pillar**, since the launch of the AHDS in 2015, AP-HP is increasingly investing each year into the development of robust IT solutions to foster collaborative data use by AP-HP both healthcare professionals and public or private partners. The IT staff in charge of the AHDS increased from 5 to +60 persons between 2015 and 2021. In order to meet a growing demand exceeding the capacity to fulfill data requests, funding approaches, such as fee-for-service, are currently under consideration. AP-HP created the AP-HP Research Foundation in 2015 to support biomedical and health research. The return on investment currently consists in publications (>20 articles in 4 years) and grants gained for large-scale national and European projects.

The **impact of the COVID-19 pandemic on the AHDS** has been deep and wide. One month after the admission to the AP-HP of the first patient diagnosed with Covid-19, a specific database was compiled from the CDW. In addition to demographic data, patient pathways, DRG codes (diagnoses and procedures), lab tests, drug prescription, the COVID database includes data of interest automatically extracted from clinical notes using NLP as well as medical images). An information notice specific to the COVID database has been added to the CDW website. The COVID database has been used to produce dashboards informing the crisis unit on a daily basis on the clinical and biological characteristics of patients diagnosed with Covid-19, the evolution of their state of health, and the admission and treating capacities of the hospitals. Evidence-based guidance has been produced for the direction of the AP-HP as well as for policymakers and health authorities on how to mitigate the impact of the outbreak. In the field of research, more than 70 data access requests have been processed by the IRB according to an accelerated procedure. The Covid-19 Research Steering Committee has also prioritized 14 therapeutic trials and 23 other studies. The standardization of data of the COVID database has enabled AP-HP to join the international consortium 4CE (Consortium for Clinical Characterization of Covid19 by EHR (4CE)) whose objective is to rapidly share and analyze aggregated data from hospitalized patients diagnosed with Covid-19 on a large scale in order to better characterize this disease [8].

4. Discussion and Conclusion

In the era of big data and machine learning, the use of RWD, and especially of EHR data will play a growing role in evidence generation in association with randomized clinical trials for promoting scientific research and improving health care delivery in the public interest. The AP-HP Health Data Space (AHDS) is a key facilitator for hospital management, clinical research and digital innovation. The AHDS has been set up with a view to a "rapid and reactive learning system", with the objective of enabling healthcare professionals to access consolidated and enriched health data, regardless of the context – care or research – in which they are collected, and to use these data to improve patient care. The current health crisis highlights the need for healthcare institutions to continue the development and deployment of CDW and Big Data spaces, to strengthen their expertise in data science and to implement efficient data quality monitoring programs. University hospitals are playing and will play a major role in real-world evidence generation. Given the level of funding dedicated to the development IT infrastructures that facilitate high-quality data access and use and the expectations placed on data-driven approaches in biomedical research and health care, there is a need of sound evaluation of the return on public investment in terms of impact of health data reuse infrastructure and projects on care processes, productivity and costs, patient safety, care quality, or health outcomes [10].

References

[1] Eichler H-G, Pignatti F, Schwarzer-Daum B, Hidalgo-Simon A, Eichler I, Arlett P, et al. Randomized Controlled Trials Versus Real World Evidence: Neither Magic Nor Myth. Clin Pharmacol Ther. 2021 May;109(5):1212–8.

[2] Rogers JR, Lee J, Zhou Z, Cheung YK, Hripcsak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. J Am Med Inform Assoc. 2020 Nov 4;28(1):144–54.

[3] Kwakkenbos L, Imran M, McCall SJ, McCord KA, Fröbert O, Hemkens LG, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. BMJ. 2021 Apr 29;373:n857.

[4] Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak. 2021 Feb 15;21(1):54.

[5] Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc. 2015 Jan;22(1):43–50.

[6] Richesson RL. Learning health systems, embedded research, and data standards-recommendations for healthcare system leaders. JAMIA Open. 2020 Dec;3(4):488–91.

[7] Griffon N, Pereira H, Djadi-Prat J, García MT, Testoni S, Cariou M, et al. Performances of a Solution to Semi-Automatically Fill eCRF with Data from the Electronic Health Record: Protocol for a Prospective Individual Participant Data Meta-Analysis. Stud Health Technol Inform. 2020 Jun 16;270:367–71.

[8] Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109.

[9] Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2019 Nov;181:104804.

[10] Campion TR, Craven CK, Dorr DA, Knosp BM. Understanding enterprise data warehouses to support clinical and translational research. J Am Med Inform Assoc. 2020 Jul 1;27(9):1352–8.