Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-16

Integrating Biobank Data into a Clinical Data Research Network: The IBCB Project

Guillaume BOUZILLE[a,1], Vianney JOUHET[b], Bruno TURLIN[a], Bruno CLEMENT[a] Mireille DESILLE[a], Christine RIOU[a], Moufid HAJJAR[b], Denis DELAMARRE[a], Danielle LE QUILLEUC[a], Frantz THIESSARD[b], Marc CUGGIA[a]

a) Inserm, Univ Rennes, CHU Rennes, Inra, Laboratoire Traitement du Signal et de l'Image (LTSI-U1099), CIC-1414, Centre de Données Cliniques, Nutrition Metabolisms and Cancer (NuMeCan), The liver biobanks network, CRB-Santé, Biosit, Biogenouest, Rennes, France.
b) Inserm, Univ Bordeaux, CHU Bordeaux, Pole de santé publique, Service d'information médicale, unit IAM, Bordeaux, France

Abstract. Development of biobanks is still hampered by difficulty to collect high quality sample annotations using patient clinical information. The IBCB project evaluated the feasibility of a nationwide clinical data research network for this purpose. Method: the infrastructure, based on eHOP and I2B2 technologies, was interfaced with the legacy IT components of 3 hospitals. The evaluation focused on the data management process and tested 5 expert queries in Hepatocarcinoma. Results: the integration of biobank data was comprehensive and easy. Five out of 5 queries were successfully performed and shown consistent results with the data sources excepted one query which required to search in unstructured data. The platform was designed to be scalable and showed that with few effort biobank data and clinical data can be integrated and leveraged between hospitals. Clinical or phenotyping concepts extraction techniques from free text could significantly improve the samples annotation with fine granularity information.

Keywords. biobank, data integration, interoperability, big data, data sharing

1. Introduction

Biobanks operate at the interface between patient care in hospitals and clinical, translational or basic researches. The cooperation and extensive collaboration through a network of multiple organizations is encouraged to enable the streamlined exchange of bio specimens and associated data. As such, biobanks are central for the development of both academic and industrial R&D, which requires an easy access to biological resources and associated data, to generate innovative drugs and biomarkers related to specific diseases [1]. With the development of high throughput genomics and big data systems, even a single experiment with human samples may give rise to huge amounts of hits, whose interest and specificity strictly depends on the quality of the original data linked to the samples. However, the development of biobanks is still hampered by the difficulty to collect and to process human bio specimens based on standards that support quality, regarding storage, phenotyping and clinical annotations including medical, genealogical, and lifestyle information in a biobank. In parallel, Clinical Data warehouse (CDW)

[1]Guillaume Bouzillé, LTSI, rue du Pr Bernard, 35043 Rennes guillaume.bouzille@univ-rennes1.fr

technologies and now Clinical Data Research Networks (CDRN) are coming forward as one of the solutions to address bio clinical data exploitation and data sharing at multiple scales. In these networks, stakeholders provide to the research community a part of their data while maintaining a data-sharing control at any time. In this context, the main objective of the iBCB project (**i**ntegrating **B**iological and **C**linical data for **B**iobank) was designed to explore, through a proof of concept, how semantic integration and CDW technologies could enrich biobanks data and facilitate sharing sparse information, that was up to now compartmentalized into clinical information systems. The aim of IBCB was to design a multi-site platform prototype capable to provide bio clinical information for biobanks in an efficient and secure way. In this paper, we present the technical infrastructure and the evaluation of the platform Hepato-Cellular Carcinoma (HCC), which is a critical research field. Indeed, Chronic liver disease incidence leading to liver cancers is increasing dramatically in the last 20 years worldwide. In France, the incidence of chronic liver disease has increased by 2.5-fold [2].

2. Material and methods:

Definition of the use case: The project involved two academic hospitals (Rennes and Bordeaux) and a Cancer Center (CLCC Bergonié of Bordeaux). To define the road map of the IBCB platform, we interviewed the potential end users (Pathologists and physicians) of the three hospitals, to identify their needs and their functional requirements regarding data reuse. We collected the functionalities and the different queries they would like to perform on the platform to conduct their research. A main requirement was to get the count of patients meeting specific clinical and biological criteria and having one or several samples, in a multi-site fashion. Five examples of queries were provided by users:

Q1: all patients having sample(s) AND with HCC AND with other non-hepatic tumor
Q2: all samples of liver tumor of patient with HCC AND with other non-hepatic tumor
Q3: all patients having liver sample(s) with HCC AND with non-cirrhotic liver
Q4: all patients having liver sample(s) with HCC AND NASH syndrome
Q5: all patients having liver sample(s) with HCC AND with cirrhotic liver and AST > 800 UI/L or ALT>800 UI/L

Infrastructure design and technical aspects: One challenge of the project was to design a scalable architecture that could be extended to several hospitals. As a proof of concept we build the architecture (fig. 1) on different CDW technologies. Two types of CDW (eHOP and I2B2) were used: eHOP is a CDW developed by the team of Rennes, which is used in 8 hospitals within the western CDRN of France [3]. I2B2 is an Open Source CDW developed by the Boston university to facilitate the use of the clinical patients' data in the translational informatics [4]. I2B2 is used in Bordeaux as a legative CDW and in Rennes to share structured data coming from eHOP. SHRINE was the third technical component that allowed to distribute query and to compute counts of patients from I2B2 endpoints. All technical components of the infrastructure were hosted within the information systems of the 3 hospitals.

Biobank and clinical data integration: The first step of data integration consisted in developing a specific ETL job to extract, transform and load sample data coming from the legacy biobank softwares (that were different in the 3 sites) in each CDW: TD biobank for Rennes and TumoroteK for Bordeaux. An ontology of data elements taking into account the comprehensive content of information available in the biobank software databases was commonly defined by the 2 hospitals: the clinical datasets included relevant structured data: ICD-10 and ICD-O diagnosis codes, Procedures Codes

(CCAM) and pathology codes (French ADICAP terminology). For lab tests, each site had its own interface terminology, so we used LOINC to map a subset of relevant data elements.

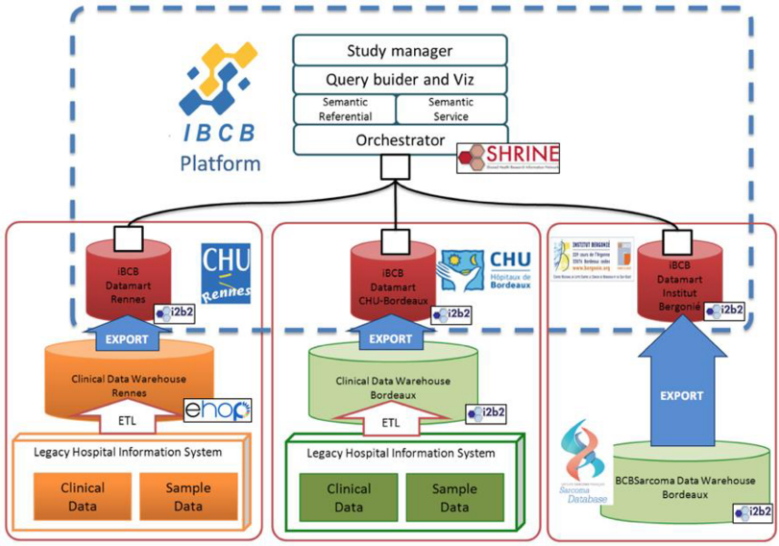

Figure 1: IBCB platform architecture

The second step was to export from the legacy CDWs, clinical and samples data in the I2B2 DataMart intended to be shared between hospitals through the SHRINE query orchestrator.

Validation and evaluation methods: Data management study: this first study was carried out to assess the ETL processes and the data integration at each level of the platform. Counts of data elements were compared from the sources to the target Datamart. Biobank data managers were solicited to evaluate the data quality before and after integration in the legacy CDW and the target Datamart. A random sample of 100 records was compared by the data managers from the two sites. Functional evaluation of the platform: A second study consisted into executing queries provided by the users to test the platform. Such queries were performed on the legacy CDW and on the IBCB datamarts. The objective was to compare the capability of each component to provide consistent and complementary information.

3. Results

Data management study: Table 1 compares from the 2 sites the count of patients and data elements integrated in the CDW:

		Rennes Site	Bordeaux Site
Available Bioclinical data collection in CDW:	- Nb of patients - Nb of bioclinical documents - Nb of related data elements - Period of time	1.2 millions 38 millions 299 millions 1995 to 2017	140,000 10 millions 235 millions 2010 to 2017
Biobank data integrated in CDW:	- Nb of samples / Nb of patient - Nb of data elements - % integrated / biobank software	33,074 / 4,958 708,323 100%	18,086 / 13,535 257,552 100%

	- Period of time	2010 to 2017	2006 to 2017
Data exported to I2B2 shared DataMart	- Nb of patients - Nb of data elements - Period of time	4,958 7,428,426 2010 to 2017	13,535 33,061,726 2006 to 2017

Functional evaluation of the platform: We performed a set of queries to compare the results from the IBCB infrastructure with those coming from the legacy CDW. Table 2 shows the results of the queries at the different stages of the platform.

Query executed on :	**Q 1**	**Q 2**	**Q 3**	**Q 4**	**Q 5**
CDW Rennes (eHOP) **CDW Bordeaux (I2B2)**	**34 patients** **84 patients**	**34 samples** **128 samples**	**84 patients** **170 patients**	**3 patients** **-**	**30 patients** **71patients**
DataMart Rennes (I2B2) **DataMart Bordeaux (I2B2)**	**34 patients** **84 patients**	**34 samples** **128 samples**	**84 patients** **170 patients**	**-** **-**	**30 patients** **71 patients**

The Figures 2 and 3 show the Biobank data representations into eHOP and I2B2 user interfaces. Specifically, eHOP enables visualization of documents and not only data elements. Integration of samples information was fully validated by data managers of biobank software.

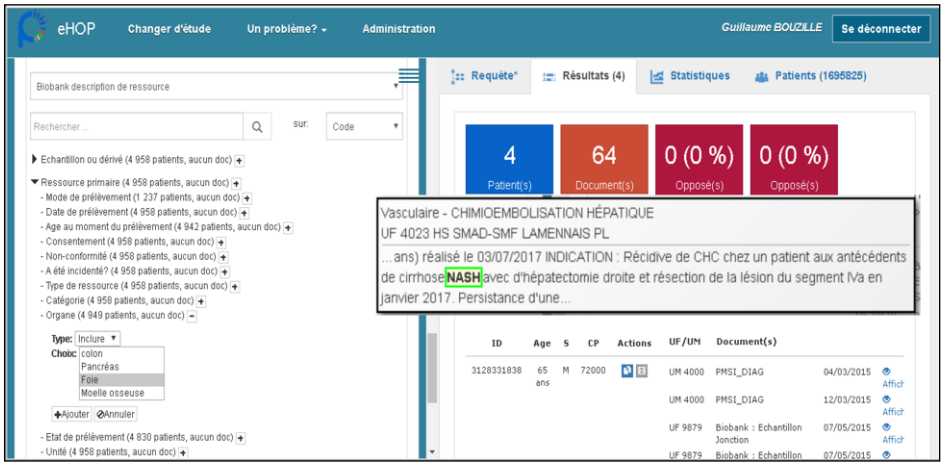

Figure 2: eHOP user interface: hierarchy of biobank data elements. Result including free text research

4. Discussion and conclusion:

The aim of the IBCB project was to investigate whether biobank data combined with bio clinical data could be shared with the researcher community at a nationwide level through a flexible and scalable infrastructure. This first attempt successfully showed that such approach is feasible and could leverage existing technologies. This needed few efforts regarding data integration, since biobanking items are quite standardized from one site to the other. The limited scope of bioclinical data used in the project also helped data integration. The collection of data elements was natively encoded with reference terminologies excepted lab tests, which required a manual mapping with the LOINC terminology.

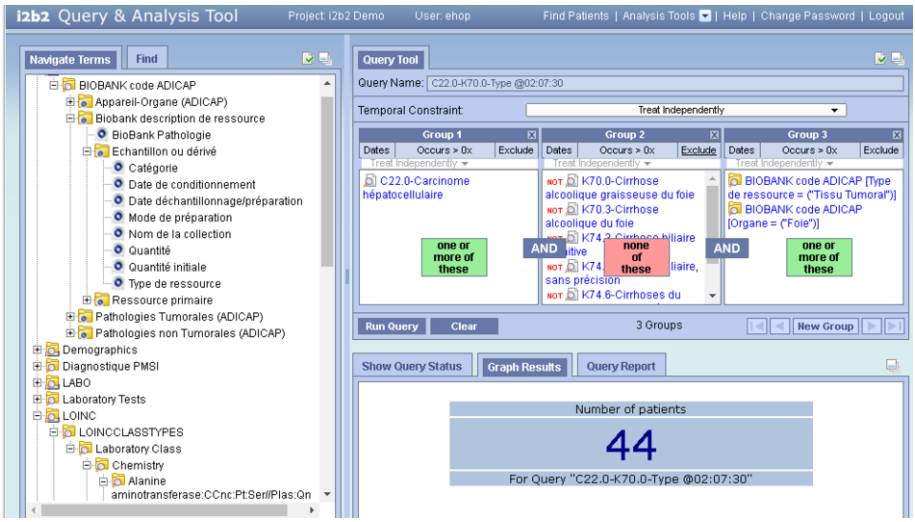

Figure 3: I2B2 User interface with hierarchy of biobank data elements

As a limit, our first experiment queried and shared data from only two sites. However, we used scalable and open source components (I2B2 and Shrine). Query 4 focused to find patient with a NASH syndrome, which turned out to be not currently coded with existing reference terminology. As only structured data was transferred to Datamart, query 4 failed on I2B2 but succeeded with eHOP (eHOP has the advantage to natively enable advanced information retrieval on both structured data and free text documents). Even if free text queries generate noise, from the user's perspective, the workload to manually review the cases returned was negligible compared to the benefit. This shows that inferring phenotypes from unstructured data is crucial to answer user needs with technologies like I2B2. Future works will focus to deploy the IBCB platform on a larger number of hospitals and to provide at a national level the existing and new services such as pre-screening functionalities, deep phenotyping [5], and data export to populate target databases such as epidemiologic registries or cohort databases.

Acknowledgements

The IBCB project was funded by the following French national infrastructures: IBiSA and BIOBANQUES

References

[1] S. Sarojini, A. Goy, A. Pecora, and K.S. Suh, Proactive Biobanking to Improve Research and Health Care, *J. Tissue Sci. Eng.* **3** (2012). doi:10.4172/2157-7552.1000116.

[2] Santé publique France. Etat de santé de la population en France : rapport 2017, (http://www.santepubliquefrance.fr/Actualites/Etat-de-sante-de-la-population-en-France-rapport-2017 (accessed February 22, 2018).

[3] D. Delamarre, G. Bouzille, K. Dalleau, D. Courtel, and M. Cuggia, Semantic integration of medication data into the EHOP Clinical Data Warehouse, *Stud. Health Technol. Inform.* **210** (2015) 702–706.

[4] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, *AMIA Annu. Symp. Proc. AMIA Symp.* (2006) 1040.

[5] W.-Q. Wei, and J.C. Denny, Extracting research-quality phenotypes from electronic health records to support precision medicine, *Genome Med.* **7** (2015) 41.