MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-303

Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach

Guillaume Bouzillé[a,b,c,d], Richard Westerlynck[d,k], Gautier Defossez[f], Dalel Bouslimi[g,h], Sahar Bayat[i], Christine Riou[a,b,d], Yann Busnel[j], Clara Le Guillou[h], Jean-Michel Cauvin[h], Christian Jacquelinet[j], Patrick Pladys[c], Emmanuel Oger[e], Eric Stindel[h], Pierre Ingrand[f] , Gouenou Coatrieux[g,h], Marc Cuggia[a,b,c,d]

[a] INSERM, U1099, - [b] Université de Rennes 1, LTSI, Rennes, 35000, France
[c] CIC Inserm 1414, [d] Centre de Données Clinique, [e] Équipe CTAD-PEPI, CHU de Rennes, 35000, Rennes, France.
[f] Registre général des cancers de la région Poitou-Charentes, Poitiers, France.
[g] Institut Mines-Télécom; Télécom Bretagne, [h] INSERM 650 Latim, Brest, France
[i] EHESP Rennes, Sorbonne Paris Cité, France,
[j]Agence de la biomédecine, Saint Denis, France,[k] CREST(ENSAI) / Inria Rennes, France

Abstract

Sharing and exploiting Health Big Data (HBD) allow tackling challenges: data protection/governance taking into account legal, ethical, and deontological aspects enables trust, transparent and win-win relationship between researchers, citizens, and data providers. Lack of interoperability: compartmentalized and syntactically/semantically heterogeneous data. INSHARE project using experimental proof of concept explores how recent technologies overcome such issues. Using 6 data providers, platform is designed via 3 steps to: (1) analyze use cases, needs, and requirements; (2) define data sharing governance, secure access to platform; and (3) define platform specifications. Three use cases – from 5 studies and 11 data sources – were analyzed for platform design. Governance derived from SCANNER model was adapted to data sharing. Platform architecture integrates: data repository and hosting, semantic integration services, data processing, aggregate computing, data quality and integrity monitoring, Id linking, multisource query builder, visualization and data export services, data governance, study management service and security including data watermarking.

Keywords:

Information Dissemination, Information Storage and Retrieval, Registries

Introduction

Health Big Data (HBD) is more than just a very large amount of data or a large number of data sources. It also refers to the complexity, challenges, and new opportunities presented by combined analysis of data. Health data collected or produced are now potentially sharable and reusable. They can be exploited at different levels and across different domains, especially concerning questions related to multidisciplinary research. This huge amount of data holds the promise of supporting a wide range of medical and health care functions, including among others clinical decision support, disease surveillance or population health management [1]. This explains the incentive policy of opening HBD around health data science being supported by different public authorities and scientific communities such as OpenData, AVIESAN or Inserm initiatives, as well as European research programs like IMI or Horizon 2020. Recently, strong initiatives have been launched in U.S to enhance utility of health Big Data and finally to enter in the next level of knowledge discovery [2].

In this context, clinical data warehouse (CDW) technology comes forward as one of the solutions to address HBD exploitation. CDW, are becoming increasingly widespread in U.S, being put to use for different purposes including cohort discovery, biomarker detection, feasibility studies or enrolment of patients in clinical trials. Research communities are currently connecting CDW to one another with the aim of creating Clinical Data Research Networks (e.g. PCORNET [2]) or biomedical research network (e.g. Data to Knowledge).

In these networks, data providers such as researchers, health facilities, research agencies or institutions make part of their data available to research community while maintaining data sharing control at all time. Thus, these trusted third-party platforms integrate and open-up scientific or potentially scientific health data [3]. This makes use of these data at a large-scale possible. In France, such platform, that would be able to integrate and share multisource and multiscale big and small health data produced by health institutions for research purposes, does not exist.

This is the aim of the INSHARE French national project; in which different and actual key issues like governance, organizational, and technical factors to perform such data sharing will be explored and addressed. The absolute goal is to facilitate access to data and foster collaborative research and data sharing between researchers and data providers. In this paper, we present and discuss these key issues and approach we are following to design the platform. The approach is driven by real research use cases of high interest for individuals and states.

Background

Data to share: In its large acceptance HBD sources comprise various types of data from structured information such as OMICS data, administrative or billing data, drug prescription data consisting of dates and dosages captured through standardized ePrescription system, to unstructured and textual data such as clinical narratives that describe medical reasoning behind prescriptions [4]. Beyond data generated by hospitals, several health data sources come from health registries or insurance databases, which are a valuable source of standardized, longitudinal, population-wide data. For instance, the French health reimbursement database (Système National d'Information Inter-Régimes de l'Assurance Maladie, SNIIR-AM) contains individualized, anonymous and comprehensive data for all health spending reimbursements received by affiliated subjects, including basic patient demographic data such as age, gender, medical drugs, and outpatient medical cares – prescribed or performed by health professionals from both public and private practices. SNIIR-AM is also linked via a unique personal health number to the French hospital discharge database (PMSI), which contains diagnostic codes, medical procedures, and admission dates for all hospitalizations. Data from SNIIR-AM is increasingly used for research projects, especially relating to detection of drugs' adverse effects in epidemiology or clinical research. The designed platform governance models were derived from the SCANNER model and adapted for data sharing.

Sharing barriers: The regulatory hurdles obstructing optimal use of data for research have been extensively discussed within specialised literature [5]. Identified factors are characterized by (i) over-cautious approach among data custodians, many of whom are unwilling to link or share data, (ii) legislators' failure to consider flexibility required to allow and support such linking and sharing and (iii) incorporation of 'good governance' models or intelligent design of working instances not contemplated within the regulatory framework, nor reflecting on the subject [5], [6]. Sethi proposes a model for data sharing governance including (i) guiding principles and best practices, (ii) safe, effective and proportionate governance, (iii) articulation of roles and responsibilities of data controllers and data processors and (iv) development of a training program for researchers that covers appropriate vetting procedures prior to sharing valuable data.

Cornerstone of data sharing and reuse is trust. Therefore, implementing a trustworthy process for handling citizens' and patients' health data is a pivotal goal. Based on the definition of a trusted relationship, one party (trustor) is willing to rely on the actions of another party. In addition, the trustor abandons (full) control over the actions performed by the trustee. As a consequence, a trustworthy system is that in which, the trustor can "place his/her trust and rest assured that the trust will not be betrayed". A system for data reuse should thus prove its trustworthiness by fulfilling the responsibility of dealing with data within the limits of a social contract regulated by policies between citizens and organizations handling the system.

The technological components behind a trustworthy system involve designing and implementing IT tools and services capable of guaranteeing data quality and security while providing interoperability, adaptability, and scalability. Specific projects funded by the EU and by the IMI initiative [7], such as EHR4CR, are dealing with such challenges, with the prospect of defining use cases, tools, technologies and a business model for data reuse. In particular, the EHR4CR business model includes accreditation and certification plans for EHR systems that can be integrated within a system for data reuse. The purpose of data reuse has implications that belong to the realm of policies and regulations, which are essential aspects for establishing trust. How to manage informed consent is one of the key aspects connected to this issue. In fact, current regulations in many European countries, which are similar to the US, with the HIPAA act, assume that consent (implied or explicit) for use of data is strictly limited to the purpose for which data were collected. This may seriously limit the scope of data analysis.

This theme needs to be reconsidered in the light of the existence of a proper, trustworthy system based on an agreement between citizens and healthcare organizations. Specific practical examples of policies for handling data reuse are provided by regional initiatives in Europe, two such cases being the United Kingdom and Catalonia. ISO/TS 14265:2011 provides a classification of different purposes for processing personal health information that can help make policy formulation more granular.

Methods

To design organizational and technical dimension of the INSHARE platform, an iterative and 4 step bottom-to-top approach has been adopted, by analyzing on the ground, existing needs, use cases, and actual difficulties encountered by the project partners. Four partners are involved in the project as data providers: 2 academic hospitals (CHU Rennes and Brest) which provide datamarts from their Clinical Data Warehouse (eHOP-CDW), 3 epidemiologic registries at a regional or national scale.

This approach aims at defining technical and functional specifications, data protection policies and governance for an efficient and valuable data sharing. Furthermore, this approach takes into account the fact that some technological issues have to be addressed and especially the evolution needs for data analysis and security tools in the scaling-up to HBD.

Step 1 - To describe use cases and user needs:

The aim of this step is to define precisely scenarios from an operational perspective, the information workflow and system/actor interfaces that relate to exploitation of health and research data via the INSHARE platform. Relevant scenarios leverage the richness and variability of data sources hosted by the platform in terms of domain, quality, and origin. Herein, the objective is to identify the functional needs, which are expected by different users of the platform: researchers- users, data providers, and internal operators of the platform.

Step 2 - To define data sharing governance and secure access to the platform:

Regarding ethical, legal and deontological aspects, a focus group composed of domain experts and representatives of patient associations conducts this study. According to the state-of-the-art step and specified use cases defined at the first step, the objective is to establish governance guidelines guaranteeing data protection and individuals' privacy rights. This step includes submission of these guidelines for validation to institutional and regulatory authorities such as the Comité consultatif sur le traitement de l'information en matière de recherche (CCTIRS) and the Commission Nationale de l'Informatique et des Libertés (CNIL), two French authorities in charge of such regulation aspect.

Step 3 - To define INSHARE platform specifications:

The aim of this step is to define a comprehensive description of intended purpose and environment of the platform. These specifications describe what the software does and how it will be expected to perform, taking into accounts the operational scenarios, security aspects, and stakeholders (users, data providers, and data managers) inputs. It also addresses some key issues in relation with data analysis and security. In terms of data protection in the scaling up, special interest is given to data traceability and on how to give back some control to data providers on the data they make available to researchers. On one hand, users have to know of their action accountability and, in another hand, patient or data provider consent for data exploitation duration has to be guaranteed. Database watermarking, a very recent solution [8], is one of the technology actually explored for those purposes.

Each data provider is part of the INSHARE project to bring their knowledge and experience with their respective data. Data providers are thus responsible for supplying necessary data to the platform in order to answer to the use cases. They have to supply all necessary information about data to correctly perform their integration and to subsequently give capability to the platform to provide the best-suited data for each user request.

Results

Use Cases: Three main use cases corresponding to 5 studies and application domains have been identified and chosen to be performed on the platform. Being able to ensure one of them will be of great interest for cares of individuals and populations. Table 1 illustrates for each use case the sources of data, which will be shared and used in the different INSHARE platform studies.

Table 1 – Use cases and application domain

Use Case	Study and application domain	Data Sources
Health care Trajectory analysis	- Pre & post-dialysis care trajectory of end-stage renal disease patients starting dialysis in emergency - Characterizing the healthcare trajectories of children (and their mother) included in Birth Defect Registry	Kidney Failure registry (REIN) SNIIR-AM Birth defect registry eHOP-CDW
Registry enrichment	Assessment of association between cancer incidence and diabetes in end-stage renal disease patients	Kidney Failure registry (REIN) Cancer registry SNIIR-AM
Signal detection	- Influenza surveillance - Adverse drug effect surveillance	eHOP-CDW SNIIR-AM Sentinel Network Open Data

For instance, regarding study on health care trajectories of end-stage renal disease patients, comorbidities are currently collected and registered in the REIN database at initiation of renal replacement therapy (RRT). But occurrence of comorbidities after RRT started is not a mandatory field of REIN [9]. Moreover, no information on prescribed treatments is available in the REIN database. Through the INSHARE platform, the hospital CDW (Rennes and Brest) will be used to collect comorbidities and expensive drug prescriptions while drug exposures will be extracted from SNIIR-AM in order to enrich the REIN registry with accurate comorbidities and medications including standards and dates of occurrence.

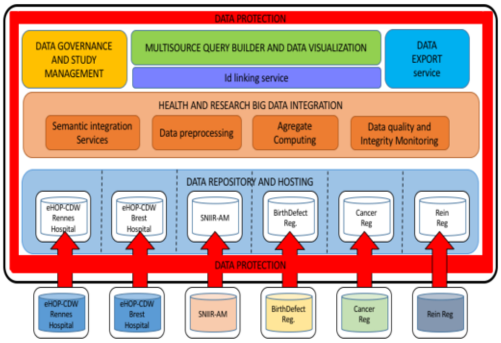

Figure 1 – Uses case under the scope of the Inshare Platform

Platform governance:

The governance determines how the range of controls and procedures with contractual obligations work together to ensure an end-to-end secure and trustfully platform, where the security and reliability of data is guaranteed. Indeed, several.
INSHARE use cases imply to perform Id linkage processing and require the access to identified data (e.g., Epidemiologic registry enrichment). Moreover, all use cases require aggregating data coming from multiscale institutions (local academic hospitals, regional to national registries, and for the SNIIR-AM, data issued from a nationwide database). Some of them imply intensive computation on big volume of data (e.g. signal detection). All these constraints have led to defining a model of governance for the platform adapted to big data sharing. The model we propose is derived and adapted from the Distributed Scalable National Network for Effectiveness Research (SCANNER). It consists of identifying 10 basic requirements: platform and data provider information, institution information, study information, ethical agreements (coming from an independent IRB-like comity), Data Sharing Agreement (from data providers who are involved in the study), approved users (external users and internal operators of the platform), authentication and access, data use, audit and accounting, patient rights, data segregation. In addition, according to the type of personal data, it includes de-identification, data watermarking, individual access, correction, openness/transparency, individual choice, use and disclosure limitation, integrity, accountability, and safeguards. Technology and some others can meet requirements by contracts, attestation of users, or management supervision.

Platform Design:

As a result, we designed the platform architecture shown in Figure 1. This architecture is oriented to meet the different use cases under the scope of the project, and to perform expected data processing while respecting governance framework mentioned above. The platform encompasses services of several weakly coupled components, the whole in a cloud-oriented architecture. Hereby, we detail some of the key components and services:

Data repository and hosting component is a buffer zone where data providers make available required datasets or datamart to share. The core idea is to host in the platform data with the finest granularity and in their most original form, i.e., with the least transformation possible.

Health big data integration layer comprises components and services dedicated to data integration and processing. The semantic integration service (SIS) contains information models (i.e. database schemas of the data sources such as eHOP or SNIIR-AM) as well as semantic resources either used in the sources or required for semantic integration (reference or interface terminologies, ontologies and mappings) and ensures standards' interoperability (such as HL7, PN13, HPRIM). SIS provides tools and methods to the other standard components.

Data preprocessing service is devoted to data enrichment. It includes NLP tasks and data indexing to make easier extraction of useful information from large-scale data stored across the different INSAHRE sources. A core functionality of these services is to execute data processing tasks and to query the data virtualization layer in order to access stored data. The developed engine is also responsible for planning, coordination, and execution of queries to the data virtualization layer in a distributed manner commercial data processing frameworks and parallel relational database management systems.

Aggregate computing service is designed for building online auxiliary indexing and summarization structures based on the incoming data processing tasks and their data requirements. Based on profiling and statistic information of the submitted processing tasks. For instance this service is used to compute from CPOE data, aggregates of drug dose per day, week, stay or globally for a patient or a population.

Data quality and integrity monitoring: The INSHARE platform deals with data sources having heterogeneous data quality, from EHRs to epidemiologic registries. Integration process has to manage such quality disparities. This component is dedicated to compute metrics to monitor data quality during the integration process. These metrics are useful to (i) alert data providers and take corrective actions at the data source, (ii) perform more accurate analysis taken into account possible bias due to data quality issues, (iii) improve data quality within the INSHARE platform, each source bringing complementary information. For instance, for the same patient, in and out hospital drug information come from different sources and is registered in different ways (structured and coded data from CPOE or SNIIR-AM, text for clinical charts, forms and notes). One source can provide more accurate or exhaustive information to others.

Id Linking service: For security reasons, in France, as in most countries, there are currently no patient identifiers that can be used to directly link data from different data sources. Nonetheless, several national programs or initiatives provide researchers either trusted third-party linkage services, or big, pre-linked datasets. This, for instance, is the case of the French hospital discharge database (used as part of the hospital billing system) that matches data coming from all hospitals in France. The SNIIR-AM is arguably one of the most noteworthy linked data sources recently opened to the research community. The Id Linking service reuses and provides methods to link data sources using deterministic and probabilistic approaches on common data elements. For instance DRG data coming from hospital are already linked with the other data of SNIIR-AM. EHOP CDW includes the DRG data. Even without specific common Id, linking can be performed using dates, groups of diagnosis and procedure codes, and ADT mode.

Data Governance, study management service and security: These services encompass tools, procedures [10] and workflow to cover governance requirements and provide continuous data protection, from their acquisition to their outsourcing and mutualization within the INSHARE platform and beyond (e.g. when exported). The idea is to complement current data protection, which mainly relies on security of the information system and which do not make it possible to know if data are used for the purposes originally foreseen, especially when data are outsourced. The protection of digital content we deploy is based on watermarking [8] and crypto- watermarking solutions (i.e. mechanisms that combine encryption and watermarking [11] that fulfill different security objectives, in particular in terms of integrity and traceability (identification of information-leak sources or of end-user misbehavior). If data are protected as long as they are not decrypted, watermarking leaves free access to them and maintains them protected by means of security attributes (e.g. digital signatures, users' ID, access rights) invisibly inserted or embedded into the data themselves. Moreover, watermarking protection is independent of the data storage format. These data protection tools are designed to take into account strong interoperability constraints so as to: i) provide security resilient to information-processing; ii) make the protection on the data provider's side compliant with the one used by the INSHARE platform and beyond.

Multisource Query Builder, Visualization and data export services: These services are intended to: design and perform complex queries on multisource data; visualize results with different modalities; and, export processed dataset to the end users. From the end user point of view, interaction with the platform consists of submitting a request for a study to the platform. Only certified and authorized operators of the platform will have to access to the query workbench for data exploitation and eventually to export required datasets to the end user.

Figures 2 and 3 illustrate the workflow for two scenarios. In the first (Figure 2), a targeted research database is fed by data extracted both from two registries (REIN and Cancer playing the role of data provider) according a study protocol about the association between kidney failure and occurrence of cancer. This protocol, which defines criteria for data selection and variables to extract from the source, and user agreement are submitted to the platform. Figure 3 illustrates how the platform is able to enrich a data source (here the Rein registry) by collecting from a list of patient ID, missing or required data (e.g. comorbidities) from the different sources. In this scenario Rein registry is user of the platform and recipient of data. All along this process, health data is maintained, secured by means of digital content protection tools, with a special interest for data traceability and audit trails.

Discussion and Conclusion

The INSHARE project's consortium made the choice to focus on some of today's crucial challenges, which, in our opinion, are still not resolved: data quality assessment for research purposes, scalability issues when integrating heterogeneous health "big data" or patient data privacy and data protection. Moreover, adoption of electronic health data is still an active

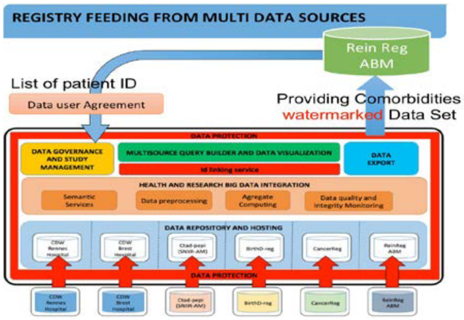

Figure 2 – Workflow for ad-hoc study and register enrichment

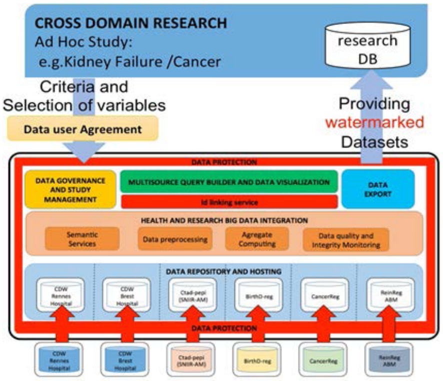

Figure 3 – Workflow for register enrichment

process and the way to exploit these data is rapidly evolving. In a second time, use of health data still requires developing to reach maturity: the aforementioned barriers have to be raised before we can obtain more significant knowledge and practical consequences. In any case, the project's use-cases were selected in view of their potentially important impact on clinical knowledge and health research, and we believe they could help demonstrate the benefits of secondary use of data. From a methodological point of view, our approach goes beyond today's most cutting-edge secondary-use technologies. We propose a combination of innovative algorithms for cryptography and watermarking supported by big data technologies, with the aim of enhancing content digital protection. Compared to other projects mainly oriented to distributed architecture, we also follow a different strategy based on a trusted third-party oriented on health-data sharing within a community of users.

Lead by the large amount of heterogeneous data available among the consortium, traditional approaches for data organization and analysis (using classical SQL server such as I2B2) are no longer efficient and quite overwhelmed. For instance, the Rennes CDW contains by itself raw data of 1.6 million patients. Recent Big Data technologies are filling the gap of this evolution that requires real time analysis, low latency, data mining, and heterogeneous unstructured data treatment. Born from needs of scalability, fault tolerance and interoperability, challenging frameworks come out as Hadoop or Cassandra. However, big data technologies are an evolving landscape. The INSHARE project is a great opportunity to test and evaluate combination of disruptive technologies and provide improved analysis performance in the perspective to meet real-world use cases and users needs, on real and massive data. For instance, preliminary tests on adverse drug effect detection have been carried out. In this example, the combination of OrientDB (which is a graph oriented database) with SPARK for aggregate computing has shown promising performance for intensive computing. Nonetheless, applying existent solutions should not be sufficient in the background of the INSHARE project. Indeed, starting from the available massive datasets, a second objective aims to design innovative algorithms and techniques in a prospective way (using a data sciences approach, sharing the statistical and computer sciences skills

Acknowledgements

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

References

[1] W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: promise and potential, *Health Inf Sci Syst*, **2**(1) (2014), 3.

[2] Patient-Centered Outcomes Research Institute (PCORI), PCORnet | National Patient-Centered Clinical Research Network. [Online]. Available: http://www.pcornet.org/. [Accessed: 23-Apr-2015].

[3] M. D. Natter, J. Quan, D.M. Ortiz et al., An i2b2-based, generalizable, open source, self-scaling chronic disease registry, *JAMIA*, **20**(1) (2013), 172-179.

[4] P. B. Jensen, L. J. Jensen, and S. Brunak, Mining electronic health records: towards better research applications and clinical care, *Nat Rev Genet*, **13**(6) (2012), 395-405.

[5] J. Peto, O. Fletcher, and C. Gilham, Data protection, in- formed consent, and research, *BMJ*, **328**(7447) (2004), 1029-1030.

[6] N. Sethi and G. T. Laurie, Delivering proportionate governance in the era of eHealth, *Med Law Int*, **13**(2-3) (2013), 168-204.

[7] M. Goldman, The Innovative Medicines Initiative: a European response to the innovation challenge, *Clin Pharmacol. Ther*, **91**(3) (2012), 418-425.

[8] J. Franco-Contreras and G. Coatrieux, Robust watermarking of relational databases with ontology-guided distortion control, *IEEE Trans. Inf. Forensics Secur*, **10**(9) (2015), 1939-1952.

[9] C. Couchoud, B. Stengel, P. Landais et al., The renal epidemiology and information network (REIN): a new registry for end-stage renal disease in France, *Nephrol Dial Transplant*, **21**(2) (2006), 411-418.

[10] G. Bouzillé, E. Sylvestre, B. Campillo-Gimenez et al., An integrated workflow for secondary use of patient data for clinical research, *Stud Health Technol Inform*, **216**, (2014), 913.

[11] D. Bouslimi and G. Coatrieux, A crypto-watermarking system for ensuring reliability control and traceability of medical images, *Signal Process Image Commun*, **47**, (2016), 160-169.

Address for correspondence

Pr Marc Cuggia,

UMR LTSI (Inserm), Faculté de médecine. Rue du Pr Léon Bernard. 35043 Rennes Cedex. marc.cuggia@univ-rennes1.fr