A De-centralized Framework for Data Sharing, Ontology Matching and Distributed Analytics

Vasileios C. Pezoulas[1], Konstantina D. Kourou[1], Themis P. Exarchos[1,2], Vassiliki Andronikou[3], Theodora Varvarigou[3], Athanasios G. Tzioufas[4], Salvatore De Vita[5], and Dimitrios I. Fotiadis[1]

[1] Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece
{bpezoulas, konstandina.kourou}@gmail.com, fotiadis@cc.uoi.gr
[2] Dept. of Informatics, Ionian University, GR 49100 Corfu, Greece
themis.exarchos@gmail.com
[3] Division of Communication, Electronic and Information Engineering, School of Electrical and Computer Engineering, National Technical University of Athens, GR 15780 Athens, Greece
{vandro, dora}@telecom.tuc.gr
[4] Dept. of Pathophysiology, Faculty of Medicine, National and Kapodistrian University of Athens, GR 15772 Athens, Greece
agtzi@med.uoa.gr
[5] Clinic of Rheumatology, Dept. of Medical and Biological Sciences, Udine University, IT 33100 Udine, Italy
salvatore.devita@asuiud.sanita.fvg.it

Abstract. In this paper we present the fundamental concept, as well as, the objectives and the architecture of HarmonicSS EU project. The proposed architecture envisages to address (i) the lack of patient stratification in large patient cohorts, (ii) the dataset heterogeneity across these cohorts, and (iii) the incomplete understanding of disease pathogenesis. In order to do so, ontology matching mechanisms are recruited for harmonizing the clinical data, distributed data analytics services have been designed for the de-centralized analysis of patient data, and different health policies can be assessed, all of them with respect to the data protection guidelines posed by the upcoming General Data Protection Regulation (GDPR). The HarmonicSS architecture has been employed towards the development of a data sharing, harmonization and distributed analytics framework for the recruitment and analysis of heterogeneous regional, national and international longitudinal cohorts of patients with primary Sjögren's syndrome (pSS). The outcomes can be used by researchers, clinicians, patients, and health policy makers.

Keywords: ontology matching, data protection, distributed data analytics, primary Sjögren's Syndrome

1 Introduction

Recent advances in medical data sharing highlight the necessity of large cohort studies in order to validate the accuracy of existing prediction models for disease prognosis,

genetic biomarkers and health policies as well. Moreover, the existing lack of patient stratification (i) increases the risk of producing incomplete results in clinical trials which employ highly expensive drugs and (ii) makes difficult the definition of evidence-based health policies. The application of existing or newly identified therapeutic targets in common clinical practice is hampered by (i) the lack of validation of potential indices in large patient cohorts, (ii) the dataset heterogeneity across these cohorts, and (iii) the incomplete understanding of disease pathogenesis. In order to fulfill these needs, data sharing, data harmonization, and data analytics are necessary.

Here, we present the HarmonicSS platform [1], a de-centralized framework that aims to address the aforementioned needs. Ethical, legal, and privacy issues for data sharing among the heterogeneous cohorts are taken into consideration in order to make this concept possible. Ontology matching tools have been recruited for overcoming the cohorts' heterogeneity and distributed learning environments have been adopted for secure processing of the cohort data. The proposed architecture is used to develop a straightforward data sharing, harmonization and distributed analytics framework for the analysis of heterogeneous regional, national and international longitudinal cohorts of patients with primary Sjögren's syndrome (pSS). Since pSS is relevant not only due to its clinical impact but also as one of the few diseases to link autoimmunity, and cancer development, its examination can establish research in many areas of medicine.

2 Towards a federated architecture

The major modules towards the HarmonicSS vision are clearly depicted in Figure 1 along with the users of the platform. These modules include the data sharing, cohort data harmonization, data sharing management, workflow, and distributed data analytics, which are further described in the sequel. The HarmonicSS platform will offer a variety of services, including big data mining tools for lymphoma prediction and patient stratification, tools for analyzing genetic data, evaluation of health policy scenarios, social media analytics, a training tool used for educational purposes by clinicians and patients, a patient selection tool for multinational clinical trials, among others.

The proposed architecture has been adopted towards the harmonization and analysis of clinical data across 22 cohorts (approx. 7500 patients in total) including a variety of pSS clinical data, such as saliva, blood and tissue samples, serum, biopsies, tears, DNA, and RNA samples. In short, pSS is a chronic inflammatory autoimmune disease causing salivary gland dysfunction (dryness in the eyes, mouth, skin, vagina), affecting primarily women near the menopausal age (almost 10 females per 1 male, followed by systemic sclerosis, systemic lupus erythematosus) [2, 3]. In 40-60% of pSS patients, extraglandular involvement is also exhibited [3], whereas 5% of pSS patients are associated with the development of B-cell non-Hodgkin lymphoma [2]. These findings support the potential of several histopathologic, cellular and molecular indices to serve as biomarkers for the classification of SS patients in distinct groups, the prognosis of disease severity and lymphoma development, and the selection of appropriate treatment. Each cohort must fulfill all the necessary data protection requirements prior to the final upload of the clinical data into the HarmonicSS platform.

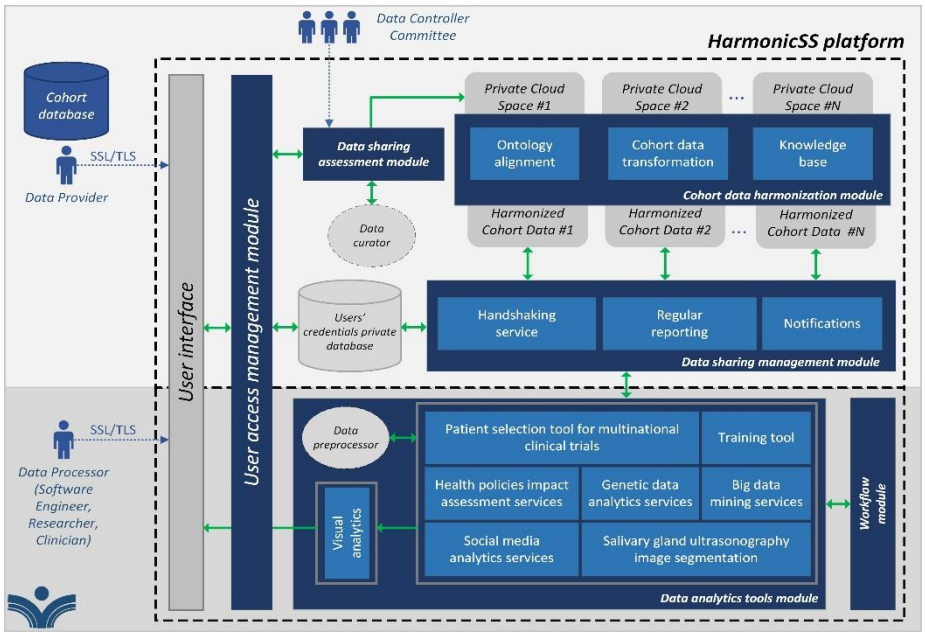

Fig. 1. The major architectural modules towards the HarmonicSS vision.

2.1 User access management

An OAuth 2 authorization framework has been adopted for user authentication and secure transfer of all the services within the platform based on Secure Sockets Layer (SSL)/Transport Layer Security (TLS) encryption protocols. Asymmetric key encryption algorithms have been implemented for the secure transfer of the clinical data into the private cloud spaces that lie within the cloud. The cloud is the environment where the HarmonicSS vision takes place. Six Kernel-based Virtual Machines (KVMs) have been already assigned for the HarmonicSS services.

2.2 Data sharing assessment

Data sharing comprises the core of the proposed federated platform. The data sharing framework is composed of two major modules, namely the data sharing assessment and the data sharing management modules. Both of them have been designed to enhance the secure evaluation and processing of the patient data with respect to the patients' privacy according to the General Data Protection Regulation (GDPR). The data sharing assessment module has been designed to ensure that data sharing fulfills all the necessary GDPR requirements from the data provider's point of view whereas the data sharing management module supervises all the processing procedures that take place within the platform from the data processor's point of view. Data providers (controllers) are responsible for providing all the required documents (e.g., signed informed consent forms, purposes of processing, transfer to third countries, data protection guarantees,

legitimate interests). The latter is conducted by a three-member data controller committee (DCC) with expertise on the pSS domain knowledge. The clinical data are finally stored on a private cloud space that is specifically designed for each cohort. Finally, a data curator mechanism is also provided for outlier detection and removal, de-duplication, attribute identification and attribute grouping, missing values detection and automatic fill based on a variety of data imputation methods.

2.3 De-centralized data harmonization

Data harmonization aims to overcome the heterogeneity of medical cohorts worldwide by converting the heterogeneous datasets into compatible ones with minimum loss. Harmonization involves several mechanisms including, cohort data transformation, terminology description, and ontology alignment. The main idea behind medical data harmonization is based on the mapping of each template of interest into a (pre-defined) reference template (model). The HarmonicSS reference model has been co-designed with the assistance of clinicians to meet the minimum requirements for effectively describing the pSS domain knowledge. Protégé [4] has been used to convert this reference model to a database ontology which comprises the main ontology. Semantic interlinking mechanisms are then used to extract the structure and the vocabulary of each source dataset for enabling ontology matching. Ontology matching is conducted by mapping the ontology of a source dataset into the main ontology (which serves as the target ontology) using semantic matching tools [5, 6]. The semantic matching algorithm provides all possible relational associations between each term of the source ontology with those from the target ontology. Then, the most appropriate terms are selected according to clinical guidelines. Novel software tools that are often employed for data harmonization include the SORTA tool [6] for data re-coding and annotation, the S-Match tool [7] for overcoming the semantic interoperability problem, the BiobankConnect software [8] for ontological and lexical indexing, and the Opal software which has been developed under the EU BioSHaRE project [9]. The harmonized data will be finally stored on the private cloud spaces of the corresponding cohorts.

2.4 Data sharing management

A data processor who wishes to process one or more de-centralized clinical cohort data (e.g., evaluate an existing or a new lymphoma prediction model on different cohorts) must first request access to the private cloud space of the corresponding cohort(s). The data providers that manage these private repositories can either provide the green light for allowing the local processing (handshaking) or not. If so, the services are executed locally on the private cloud spaces and the results are combined according to the distributed learning concept based on which the clinical data never leave the hospital (clinical center). This enhances the secure processing of the clinical data but often introduces biases during the analysis due to the heterogeneity of the cohort data structures. Harmonization is a promising solution that overcomes this heterogeneity as mentioned before. The data providers are also responsible for providing regular reports about their data management status (e.g., any actions performed by the data processors on the patients' data). Notifications can also be sent between the users of the platform.

2.5 Workflow

The workflow module is the basis for executing the data analytics services. Workflows are organized in a specific manner which includes input(s), processing stages and output(s). A workflow mechanism consists of the workflow builder which allows the user to build a model, the workflow loader which allows the user to load the constructed model and the workflow manager which allows the user to edit, execute or delete existing workflows. The user can access the workflow mechanism through the web client. For example, the workflow module allows the user to upload an existing prediction model and evaluate it on the HarmonicSS clinical cohort data.

2.6 Distributed data analytics

Data analytics consists of a variety of services including genetic data analytics, big data mining services, health policies impact assessment, social media analytics, patient selection tool for clinical trials, training tool for educational purposes. Genetic data analytics involves the identification or confirmation of existing SNPs, correlation of existing SNPS with disease sub-phenotypes, case-control and case-case associations, among others. The big data mining services consist of data mining algorithms for constructing data models (e.g., for lymphomagenesis prediction) which can be either executed on an individual cohort's private cloud space (one site analysis using classic algorithms, such as, support vector machines, logistic regression, decision trees, neural networks, etc.) or on multiple private cloud spaces using distributed data mining approaches (multisite analysis) [10] for security purposes. The health policies impact assessment procedure involves the creation, modeling, and assessment of new health policy scenarios. These models can be used to propose a drug to be tested on an appropriately selected patient and improve the patient diagnosis and screening procedure.

3 Discussion

The HarmonicSS infrastructure follows European guidelines for the management of pSS patients in order to derive rules for storing blood, tissue, saliva, serum, DNA, RNA and biopsies samples in biobanks for pSS. Prior to the data harmonization procedure, it is important to identify missing information from the cohorts. HarmonicSS has already defined a set of minimum criteria necessary at diagnosis/follow-up including pathologic ocular involvement disease indicators, such as, Schrimer's, and laboratory measures, such as, leukopenia, serum, cryoglobulinemia, tears, lip or parotid biopsy, ESSPRI and ESSDAI scores, etc., for improving data inclusion and quality as well.

Since the HarmonicSS cohorts contain approximately 500 patients that have progressed to lymphomagenesis, the risk prediction of lymphoma development will be effectively addressed. Correlation analysis between harmonized genetic phenotypes can possibly lead to the identification of new biomarkers and/or validation of existing ones, with greater accuracy. Since no data pooling is performed but rather a decentralized approach for harmonization and analytics is adopted, the security of the patient data is well-preserved and thus the platform's reliability and impact is greatly enhanced. The

HarmonicSS 'data protection by design' (GDPR compliant) architecture along with the distributed data processing services through appropriate requests, comprise a novel strategy for the construction of an innovative federated health platform for pSS patients which further considers for biobank creation and maintenance as well.

The HarmonicSS framework details a novel methodology for the initialization, maintenance, and expansion of heterogeneous data infrastructures. The overall importance and major contribution of HarmonicSS lies in the fact that it is performed on a cross-country data infrastructure of unique heterogeneity and content variability. Great emphasis is given on the development of semantic interlinking tools for homogenizing the clinical data, as well as, on the establishment of distributed environments for analyzing the harmonized data. Results linking evidence to practice can be used to support policy makers, and health management researchers to define novel health recommendations for pSS prevention to be adapted to a shared health policy environment.

Acknowledgement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731944 and from the Swiss State Secretariat for Education, Research and Innovation SERI under grant agreement 16.0210

References

1. HarmonicSS: HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making, http://harmonicss.eu/.
2. Mavragani, C, and Moutsopoulos, H.: Sjögren syndrome. CMAJ, 186(15), E579-586 (2014).
3. Ramon-Casals, M., Brito-Zerón, P., Sisó-Almirall, A., and Tzioufas, A. G: Topical and systemic medications for the treatment of primary Sjögren's syndrome. Nat. Rev. Rheum., 8, 399-411 (2012).
4. Protégé: a free, open-source ontology editor and framework for building intelligent systems, http://protege.stanford.edu/
5. Christophides, V., Efthymiou, V., and Stefanidis, K.: Entity resolution in the web of data. Synthesis Lectures on the Semantic Web, 5(3), 1-122 (2015).
6. Pang, C., Sollie, A., Sijtsma, A., Hendriksen, D., Charbon, B., de Haan, M., de Boer, T., Kelpin, F., Jetten, J., and van der Velde, J. K.: SORTA: a system for ontology-based recoding and technical annotation of biomedical phenotype data. Database, 2015 (2015).
7. Giunchiglia, F., Autayeu, A., and Pane, J.: S-Match: an open source framework for matching lightweight ontologies. Semantic Web, 3, 307-317 (2012).
8. Pang, C., Hendriksen, D., Dijkstra, M., van der Velde, K. J., Kuiper, J., Hillege, H. L., and Swertz, M. A.: BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J. Amer. Med. Assoc., 22, 65-75 (2014).
9. Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kvaloy, K., Hillege, H. L., Tasse, A. M., Ferretti, V., and Fortier, I.: Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg. Themes Epidemiol., 10, 12 (2013).
10. Tsoumakas, G., and Vlahavas, I.: Distributed data mining. Encyclopedia of data warehousing and mining (2009).