Editorial

The clinical and technical impact of the HarmonicSS project

A.V. Goules[1], T.P. Exarchos[2], D.I. Fotiadis[3], A.G. Tzioufas[1]

[1]*Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens;*
[2]*Department of Informatics, Ionian University, Corfu;*
[3]*Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Greece.*

Andreas V. Goules, MD
Themis P. Exarchos, PhD
Dimitris I. Fotiadis, PhD
Athanasios G. Tzioufas, MD, PhD

Please address correspondence to:
Andreas V. Goules,
Department of Pathophysiology,
School of Medicine,
National and Kapodistrian
University of Athens,
Mikras Asias Str. 75,
11527 Athens, Greece.
E-mail: agoules@med.uoa.gr

Received on June 3, 2021 and accepted on July 9, 2021.

Clin Exp Rheumatol 2021; 39 (Suppl. xxx): S00-S00.

Key words: Sjögren's syndrome, HarmonicSS, harmonisation, lymphoma

Competing interests: none declared.

HARMONIsation and integrative analysis of regional, national and international Cohorts on primary Sjögren's syndrome (HarmonicSS) is a European Union funded project focused on clinical unmet needs of primary Sjögren's syndrome (pSS). The main goal of HarmonicSS was to bring together the most important clinical centres specialised in SS across Europe and gather the maximum quantitative and qualitive clinical data in order to address clinically significant questions related to SS. To accomplish such an ambitious vision, distinct scientific fields were engaged and the project consisted of 21 clinical and 13 technical partners. Therefore, HarmonicSS was designed and developed in both clinical and technical level with common denominator the harmonisation process which represents the core of the project (1). The harmonisation process aims to match and transform the heterogeneous terminology used by different clinical partners for the various aspects of the disease, into a common set of terms designated as reference model. The common "scientific language" of reference model offers the opportunity to align clinical information related to SS, ensuring an automated process which homogenize and thus "harmonise" clinical data. This initial step is absolutely necessary in order to handle and utilise SS related data in an optimal way, given that the common reference model can be further expanded and enriched, reflecting the "technical" plasticity provided by the infrastructure which hosts the HarmonicSS project. Apart from the harmonisation process itself, technical challenges such as data curation, federated analysis as well as legal issues related to GDPR and data sharing, have been achieved within the HarmonicSS project.

The clinical aspects of the project are summarised as follows: a) to define distinct clinical phenotypes of the disease along with the prevalence of significant clinical, laboratory and histologic parameters in a totally harmonised cohort through integration analysis; b) to explore potential associations between the histopathology of labial minor salivary gland (LMSG) biopsy and the clinical phenotype of the disease; c) to validate old and discover novel biomarkers; d) to study and develop lymphoma classification and prediction models; and e) to employ data driven approaches contributing to the transition into the era of precision medicine. To successfully complete the clinical tasks, research activity was simultaneously developed in 3 levels, since it was not feasible to address complex scientific questions such as discovery of novel biomarkers requiring biologic samples or lymphoma classification models, by simple integration analysis. The first level of research accounts for data collection and analysis from single centres, the second level from spontaneous collaborations of more than one centres while the third level from the whole consortium, meaning all clinical partners/centres. Needless to say, that there are pros and cons for each level of analysis: the third level analysis is characterised by high statistical power but less quality of data (*e.g.*, high rates of missing values, restrictions in matching procedure between study and control groups etc.) and limited number of involved variables as defined by the reference model. On the contrary, first and second level analysis characteristics include lower statistical power, better quality of data and more explored variables.

Major progress and innovation were achieved from the technical point of view, towards the development of data governance mechanisms and data analytics services. More specifically, the data sharing assessment module provides functionalities for the upload of legal and ethical documents, the evaluation of GDPR compliance of these

documents and the subsequent application of beyond the state-of-the-art data curator mechanisms to enhance the quality of clinical data in terms of accuracy, relevance and completeness (2). The data sharing management module handles data access to the private cloud space of each data provider through appropriate notification mechanisms that ensure the transparency during processing of sensitive data. The cohort data harmonisation module provides functionalities for aligning the heterogeneous datasets using ontology-based mechanisms to enable semantic matching. The data mining services include several tools, which can support both local and federated learning scenarios, including functionalities for feature selection, data discretisation, class imbalance handling, hyperparameter optimisation and performance evaluation purposes, along with additional visualisation components for model visualisation and regression modelling. These functionalities have been used for clinical scenarios in order to address the clinical unmet needs of pSS, such as, development of lymphomagenesis models, discovery of new biomarkers and/or validation of the existing ones. The genetic data analytics services module offers functionalities for mining association rules with pre-defined support and confidence intervals across genetic datasets towards the discovery of associations between clinical sub phenotypes and single nucleotide polymorphisms (SNPs) (3). The visual analytics module provides straightforward tools for extracting hidden patterns within the cohort data through the implementation of high-performance visualisation methods, such as, heatmaps, map plots, line plots, point plots, etc. The social media analytics services module offers a single-point access to pSS-related social media posts and related content with filtering options. The health policies impact assessment services module enables the evaluation of user-defined health policy scenarios by assessing whether health impact and cost of scenarios are positive or not in the existing healthcare systems, financial figures and society, using various health outcomes (*e.g.*, patient history) as input. The patient selection tool for multinational clinical trials provides functionalities for the targeted selection of patients across the harmonised cohort data for multinational clinical trials, given a specific set of pre-defined criteria. The salivary gland ultrasonography image segmentation module applies deep learning algorithms to distil knowledge from salivary gland ultrasonography images towards the automated segmentation of the salivary gland and the classification of salivary gland ultrasonography images according to a pre-defined scoring system. Finally, the training tool provides educational material to both non-clinical and clinical experts including text, image or video, in a user-friendly environment to promote interactivity (4, 5).

Of high importance were also the results coming from application of health policy and process evaluation. Findings from survey data show variations in access, volumes of treatments delivered to pSS patients and also their perceived quality of life and satisfaction for SS care across Europe. General practitioners (GPs) play a crucial role in recognizing early SS related symptoms, in order to refer patients for diagnosis and allow in this way access to treatments for relieving symptoms and minimising complications. However, diversity of disease manifestations, makes diagnosis difficult and thus an informative tool for GPs to increase awareness in primary healthcare, focusing on methods to recognise and relate SS symptoms, is of clinical significance (4). Furthermore, given the high impact of the disease on patients' psychological and emotional status, the emerging recommendation to physicians is giving enough importance to this aspect, providing clear and complete information about patients' medical condition. In fact, such an easy intervention has a significantly positive impact on the psycho-emotional wellbeing of the patient. Despite the non-random nature of the sample, these findings are valuable in obtaining patents' perspectives and views, contributing to healthcare providers' decision and organization of healthcare delivery, tailored according to patients' status and preferences.

Moreover, analysis of population data highlights that:

- There is a need to define a valid protocol for SS patients' selection from primary care charts, facilitating population-based research (which is still scarce in the literature);
- in community setting the disease is not always confirmed through report by specialists and/or validated diagnostic criteria;
- there is a need for future investigation of novel findings from population-based analysis regarding comorbidities, use of healthcare services and drug consumption.

Ethical issues, especially when dealing with sensitive health data, are always important to consider during data sharing, harmonisation and analytics. Depending on the origin of the data, the relevant rules and regulations should be followed, *i.e.*, GPDR when dealing with EU patients' data. Within HarmonicSS, specific tools have been developed, the data sharing management and the data sharing assessment, in order to secure that ethical issues are properly handled and GDPR rules are followed. The data sharing management tool, prompts the relevant data providers to upload the study certification documents, the ethical approvals, the GDPR documents etc., which are all reviewed by the Data Controllers Committee. On the other hand, the data sharing management tools works as a handshaking procedure; the interested clinical centre wishing to obtain access to data of other cohorts, needs to have the approval (handshaking) with the cohort(s) controllers providing the data. Furthermore, the distributed/federated learning architecture provides additional safeguards when it comes to ethical issues, since the cohort data do not "leave" their premises, rather than learning the model development in a distributed manner and the algorithms go to the data, rather that the opposite.

Overall, the HarmonicSS project after conducting extensive research at the 3 levels of analysis, successfully addressed the previously described unmet needs of SS:

i) the prevalence of major and commonly used features in clinical practice

were estimated in 7,551 retrospectively harmonised patients and the effect of gender, early and late disease onset, presence of cryoglobulins, absolute seronegativity, geolocation and heavy inflammatory infiltration within LMSG, on the clinical expression of the disease was studied;

ii) the clinical spectrum of specific subsets of pSS patients was extensively investigated including males, cryoglobulinaemic and lymphoma patients, early and late disease onset patients and those with high focus score (FS) (6-10);

iii) cryoglobulinaemia, total ESSDAI score at SS diagnosis, salivary gland enlargement, rheumatoid factors and male gender were identified as risk factors associated with lymphoma after constructing novel and data driven classification and prediction lymphoma models, creating a risk stratification landscape for lymphoproliferative disorders in SS (11);

iv) older biomarkers such as CXCL13 or traditional lymphoma predictors were validated at least to some extent and new biomarkers were discovered including miRNA200b-5p in MSG biopsy specimens and serum tissue lymphopoietin serum protein (TSLP) (12-14);

v) the clinical significance of health policies at the level of primary health care for early diagnosis of pSS was highlighted;

vi) advanced and novel research, service and data analysis tools oriented to the clinical unmet needs of pSS were developed (15).

The post-HarmonicSS era refers to the period after the end of the program, aiming to retain and upgrade the infrastructure which was built during the HarmonicSS, offering the opportunity to store, enter, share and improve the quality of the cumulative clinical data hosted in the platform for continuous and advanced level of research activity. To this end, all the necessary legal and administrative issues have been included in a solid sustainability plan.

The HarmonicSS project offers clinically important perspectives far beyond the narrow field of SS, encompassing any complex and systemic disease: 1) The reference model and the harmonization process have been proven fundamental tools to describe any complex systemic disease such as systemic autoimmune diseases or diabetes mellitus, with SS being the prototype example, not only for research purposes but also for future construction of medical files in various health systems, 2) The introduction of novel, hybrid data driven approaches such as the Fast Correlation Based Feature selection (FCBF)/ binary multivariable logistic regression for integrated type of analysis, can be utilised to design classification and prediction models for diseases, instead of the typical backward, forward or stepwise logistic regression models used so far, and 3) the advanced and high level of the infrastructure which includes a variety of tools and services along with not only integration but also federated type of analysis, and above all the plasticity to adjust to future needs, may cover a wide spectrum of systemic diseases in terms of research and therefore can be adopted also by other fields.

References

1. GOULES AV, EXARCHOS TP, PEZOULAS VC *et al*.: Sjogren's syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts. *Clin Exp Rheumatol* 2019; 37 (Suppl. 118): S175-84.
2. PEZOULAS VC, KOUROU KD, KALATZIS F *et al*.: Medical data quality assessment: On the development of an automated framework for medical data curation. Comput Biol Med 2019; 107: 270-83.
3. KOUROU KD, PEZOULAS VC, GEORGA EI *et al*.: Predicting Lymphoma Development by Exploiting Genetic Variants and Clinical Findings in a Machine Learning-Based Methodology With Ensemble Classifiers in a Cohort of Sjögren's. *IEEE Open J Eng Med Biol* 2020; 1: 49-56.
4. CHATZAKI C, GOULES A, DE VITA S *et al*.: A Training Tool to support the management and diagnosis of Sjogren's syndrome. *Clin Exp Rheumatol* 2020; 38 (Suppl. 126): S174-9.
5. VUKICEVIC AM, RADOVIC M, ZABOTTI A *et al*.: Deep learning segmentation of Primary Sjogren's syndrome affected salivary glands from ultrasonography images. *Comput Biol Med* 2021; 129: 104154.
6. ARGYROPOULOU OD, PEZOULAS V, CHATZIS L *et al*.: Cryoglobulinemic vasculitis in primary Sjogren's Syndrome: Clinical presentation, association with lymphoma and comparison with Hepatitis C-related disease. *Semin Arthritis Rheum* 2020; 50: 846-53.
7. CHATZIS L, GOULES AV, PEZOULAS V *et al*.: A biomarker for lymphoma development in Sjogren's syndrome: Salivary gland focus score. *J Autoimmun* 2021; 121: 102648.
8. CHATZIS L, PEZOULAS VC, FERRO F *et al*.: Sjogren's Syndrome: The Clinical Spectrum of Male Patients. *J Clin Med* 2020; 9: 2620.
9. CHATZIS L, PEZOULAS V, GOULES AV *et al*.: Sjogren's syndrome associated lymphomas: clinical description and 10-year survival. *Ann Rheum Dis* 2021; 80 (Suppl. 1): 180-1.
10. GOULES AV, ARGYROPOULOU OD, PEZOULAS VC *et al*.: Primary Sjogren's Syndrome of Early and Late Onset: Distinct Clinical Phenotypes and Lymphoma Development. *Front Immunol* 2020; 11: 594096.
11. CHATZIS L, PETZOULAS V, GOULES AV *et al*.: Predicting risk factors of MALT lymphoma in Sjogren's syndrome. *Ann Rheum Dis* 2021; 80 (Suppl. 1): 178-9.
12. GANDOLFO S, BULFONI M, FABRO C *et al*.: Thymic stromal lymphopoietin expression from benign lymphoproliferation to malignant B-cell lymphoma in primary Sjogren's syndrome. *Clin Exp Rheumatol* 2019; 37 (Suppl. 118): S55-64.
13. GANDOLFO S, FABRO C, KAPSOGEORGOU E *et al*.: Validation of thymic stromal lymphopoietin as a biomarker of primary Sjogren's syndrome and related lymphoproliferation: results in independent cohorts. *Clin Exp Rheumatol* 2020; 38 (Suppl. 126): S189-94.
14. KAPSOGEORGOU EK, PAPAGEORGIOU A, PROTOGEROU AD, VOULGARELIS M, TZIOUFAS AG: Low miR200b-5p levels in minor salivary glands: a novel molecular marker predicting lymphoma development in patients with Sjogren's syndrome. *Ann Rheum Dis* 2018; 77: 1200-7.
15. PEZOULAS VC, KOUROU KD, KALATZIS F *et al*.: Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning. *IEEE Open J Eng Med Biol* 2020; 1: 83-90.