Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250381

Gemelli SLE Data Mart: A Real World Evidence Framework for the Evolution of Systemic Lupus Erythematousus Patients

Laura ANTENUCCI [a,b], Livia LILLI[a,b,1], Silvia Laura BOSELLO[a], Stefano PATARNELLO[a], Augusta ORTOLAN[a], Jacopo LENKOWICZ[a], Marco GORINI[c], Gabriella CASTELLINO[c], Alfredo CESARIO[a], Maria Antonietta D'AGOSTINO[a], and Carlotta MASCIOCCHI[a]
[a]*Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy*
[b]*Catholic University of the Sacred Heart, Rome, Italy*
[c]*AstraZeneca Italy, MIND, Milan, Italy*
ORCiD ID: Livia Lilli https://orcid.org/my-orcid?orcid=0009-0005-3319-7211 Silvia Laura Bosello https://orcid.org/0000-0002-4837-447X, Stefano Patarnello https://orcid.org/0009-0008-2765-5935, Jacopo Lenkowicz https://orcid.org/0000-0002-8366-1474, Marco Gorini https://orcid.org/0009-0008-1455-3884, Alfredo Cesario https://orcid.org/0000-0003-4687-0709, Maria Antonietta D'Agostino https://orcid.org/0000-0002-5347-0060, Carlotta Masciocchi https://orcid.org/0000-0001-6415-7267

Abstract. The diverse data sources in hospitals are crucial for real-world evidence (RWE) analysis and for developing decision support systems. However, integrating this variety of sources into standardized data collections for a specific disease, requires the development of highly personalized frameworks. In this study, we propose a RWE approach for the development of a Data Mart for the Systemic Lupus Erythematosus (SLE) at the Gemelli hospital of Rome. Our approach combines natural language processing and data mining procedures to capture information about organ involvements, activity episodes and treatments, under rules defined by physicians. The Gemelli SLE Data Mart includes 262 SLE patients with at least a hospitalization from 2012 to 2020, and an outpatient visit, with a total amount of 5962 contacts. We also developed a visualization tool to display each patient longitudinally and we performed a data-driven analysis to stratify patients into progression groups, based on their involved organs and flares. The Gemelli SLE Data Mart and its framework were useful to build up a standardized data collection, to be used for further personalized medicine applications.

Keywords. Real-Word Evidence (RWE), Real-World Data (RWD), Data Mart, Systemic Lupus Erythematosus (SLE).

1. Introduction

Real-world evidence (RWE) in healthcare refers to the clinical evidence derived from analysis of real-world data (RWD), i.e. data coming from both structured and

[1] Corresponding Author: Livia Lilli; E-mail: livia.lilli@policlinicogemelli.it.

unstructured sources, collected in different clinical settings such as hospitalizations and outpatient visits. RWE allows a representation of the disease, and it is increasingly recognised essential to optimise treatment strategies and to improve clinical decision-making [1, 2]. Previous studies proposed the development of computerized systems for the collection of RWD and the creation of clinical evidence [3]. In this paper we applied RWE in the context of Systemic Lupus Erythematosus (SLE) at the Gemelli hospital of Rome, with the aim of developing a Data Mart to be used in the longitudinal characterization of SLE patients in terms of disease trajectories. Our framework had to consider the complexity of SLE, where it's difficult to accurately characterize the evolution of the disease, cause of its multi-domain nature [4]. The big amount of information on organ involvements, activity events, and treatment patterns, often recorded as free-text during outpatient visits and hospitalizations, make difficult to data scientist to collect and standardize the data. To address these challenges, we developed a multi-step processing pipeline that integrates data mining and natural language processing (NLP) with the expert knowledge of physicians, ensuring accurate characterization of patient evolution across contacts (Figure 1). The Gemelli SLE Data Mart stores the disease evolution of the patients, with the support of a visualization dashboard and a data-driven analysis about the cohort stratification by progression. Data mining and NLP were developed using SAS (Viya and Enterprise Guide) and Python. Python Dash was used for the visualization tool, while statistical analyses were performed with R studio.

2. Methods

Data Collection (A). We selected a cohort of SLE patients with at least a hospitalization in the range of time from 2012 to 2020, and an outpatient visit. As input data for the framework, we considered the inpatient and outpatient reports, and the structured laboratory values. All patient information was pseudo-anonymized to ensure data privacy.
Structured Data Retrieval (B). Standard extraction, transformation and load (ETL) procedures have been used to capture structured data, mostly associated to hospitalization episodes, as laboratory data and ICD-9 diagnosis codes.
EHR Segmentation for Semantic Classification (C). Textual data were segmented and tagged with respect to key semantic areas: diagnosis, symptom, laboratory and therapy.
Categorization and Association of laboratory tests (D). Since most patients undergo laboratory tests outside the hospital, we used text mining techniques to extract these values from the EHRs and convert them into structured data. After standardizing units of measurement and categorizing values as either normal or out of range, this data was integrated with the structured laboratory information already available from the hospital's system. As final step, we linked each patient contact with the most recent laboratory information within 60 days or during hospitalization.
Clinical features identification through NLP (E). Starting from the segmented texts and the ontology defined by the clinical team, we developed an NLP pipeline in order to extract concepts related to organ domains, symptoms, and therapies. For further details the study was conducted and validated in [5].
Rule-based engines to classify Organ Domains Involvement, Activity Events and Therapies (F). The clinical team developed a system of rules, to characterize SLE patients in terms of involved organ domains, activity events and treatments [6]. Domain

involvements derive from the patient diagnosis and persist over time, while activity events were computed by combining symptoms and laboratory values at each contact. Both the organ involvements and the activity events were categorized in 8 distinct areas, that are: articular, cutaneous, haematological, kidney, serositis, systemic, neurological and vascular. Finally, treatments were extracted at each contact and categorized into 4 macro-groups: antimalarials, conventional and biological immunosuppressants, and glucocorticoids with the specification of high dosages.

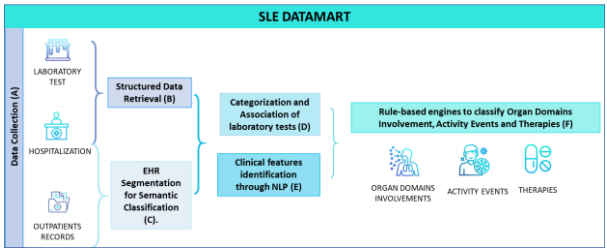

Figure 1. Multi-step processing pipeline for the Gemelli SLE Data Mart development

3. Results

The Gemelli SLE Data Mart describes the clinical history of 262 patients (88% female, median age 43) with at least a SLE admission between 2012 and 2020 and an outpatient visit, at Gemelli Hospital, for a total of 5962 contacts. Figure 2 illustrates the dashboard visualization tool, which provides a longitudinal overview of each patient.

Figure 2. Visualization dashboard for patient's longitudinal disease evolution

A data-driven analysis was also conducted to track the evolution of SLE patients over time. Patients were categorized into progression groups as mild (MiP), moderate (MoP) and severe (SP) based on the number of involvements at baseline and active domains during their history. Patients with more organ domains affected by flare than the number of baseline involvements were classified as SP, those with at least a baseline involvement and just one flare were classified as MiP and the others as MoP. As an example, if a patient presents articular, cutaneous and vascular involvements at baseline, and articular and cutaneous flares during the longitudinal, then he belongs to MoP. Figure 3 shows a scatter representation of the 3 groups, while Table 1 provides statistics on some continuous variables (Kruskal-Wallis rank sum test has been computed)

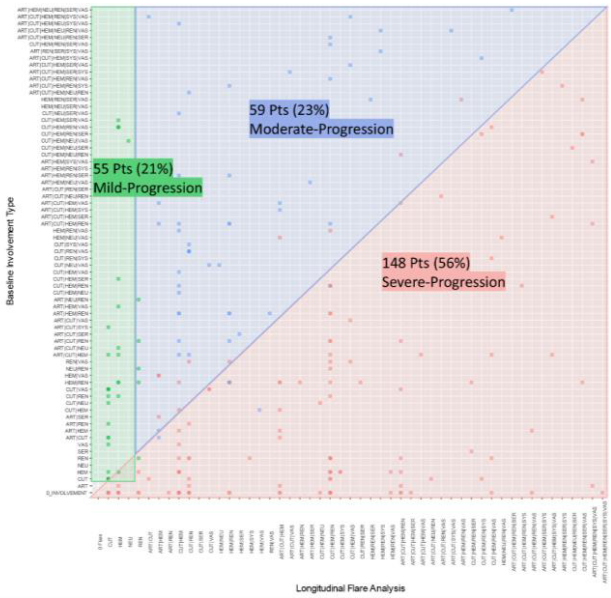

Figure 3. Data-driven analysis on SLE patients' progression over time

Table 1. Variable characteristics for the different progression group

Variable	MiP	MoP	SP	P value
Ordinary Admission	0.00 (0.00, 1.00)	1.00 (0.00, 2.00)	1.00 (1.00, 3.00)	<0.001
Contacts	9 (6, 16)	18 (12, 22)	23 (15, 35)	<0.001
Baseline Involvements	2.00 (1.00, 3.00)	4.00 (3.00, 4.50)	1.00 (0.00, 2.00)	<0.001
Longitudinal Involvements	4.00 (2.00, 5.00)	5.00 (4.00, 6.00)	5.00 (4.00, 6.00)	<0.001
Flares	1.0 (0.0, 3.0)	5.0 (3.5, 7.5)	6.0 (3.0, 11.0)	<0.001

4. Discussion

The Gemelli SLE Data Mart was used for further explorative analysis on the evolution of patients over time. Figure 2 shows the visualization tool with an upper section displaying the evolution in terms of involved organ domains in the longitudinal, a middle section showing activity events for each domain, and a lower section tracking changes in drug administration, including dosage adjustments. This tool allows physicians to explore specific patterns, such as the correlation of high disease activity with organ involvement and therapeutic decisions but is also a useful aid for making clinical decisions and highlighting trends in disease evolution. Furthermore, we developed an analytic tool for progression-based stratification of patients. Of the overall cohort, 56% resulted as SP, 21% MiP, and 23% MoP (Figure 3). The stratification into the three progression groups correlates with the increase from the mild to the severe group, in admissions, contacts, and flares, as shown in Table 1.

5. Conclusions

This work illustrates a systematic approach for constructing RWD in SLE using a multi-step framework to create a disease-specific Data Mart. This enabled data aggregation and visualisation in a dashboard in order to improve clinical decision-making by tracking patient longitudinal history and identifying disease patterns and trends at an early stage. However, the possible implementations are many, like developing predictive models to prevent severe episodes, evaluating drug efficacy, and improving early diagnosis to manage overlap and minimize missed diagnoses. The system could also be implemented in a multicentric scenario on larger cohort of patients and other chronic disease in order to test its scalability.

Acknowledgements

The project has been developed with the financial contribution of AstraZeneca and it has been approved by the Institutional Review Board of the Fondazione Policlinico Universitario Agostino Gemelli IRCCS (protocol number 0034012/22).

References

[1] Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Annals of the Rheumatic Diseases, 2023;82(3):306-311.
[2] Liu F. Data Science Methods for Real-World Evidence Generation in Real-World Data. Annu Rev Biomed Data Sci. 2024 Aug;7(1):201-224. doi: 10.1146/annurev-biodatasci-102423-113220. Epub 2024 Jul 24. PMID: 38748863.
[3] Lamer A, Saint-Dizier C, Paris N, Chazard E. Data Lake, Data Warehouse, Datamart, and Feature Store: Their Contributions to the Complete Data Reuse Pipeline. JMIR Med Inform. 2024 Jul 17;12:e54590. doi: 10.2196/54590. PMID: 39037339; PMCID: PMC11267403.
[4] Ameer MA, Chaudhry H, Mushtaq J, Khan OS, Babar M, Hashim T, Zeb S, Tariq MA, Patlolla SR, Ali J, Hashim SN, Hashim S. An Overview of Systemic Lupus Erythematosus (SLE) Pathogenesis, Classification, and Management. Cureus. 2022 Oct 15;14(10):e30330. doi: 10.7759/cureus.30330. PMID: 36407159; PMCID: PMC9662848.
[5] Lilli L, Bosello SL, Antenucci L, Patarnello S, Ortolan A, Lenkowicz J, Gorini M, Castellino G, Cesario A, D'Agostino MA, Masciocchi C. A Comprehensive Natural Language Processing Pipeline for the Chronic Lupus Disease. Stud Health Technol Inform. 2024 Aug 22;316:909-913. doi: 10.3233/SHTI240559. PMID: 39176940.
[6] Ortolan A, Lilli L, Bosello SL, et al POS1142 DEVELOPMENT AND VALIDATION OF A RULE-BASED FRAMEWORK FOR AUTOMATED IDENTIFICATION OF LONGITUDINAL CLINICAL FEATURES ABOUT SYSTEMIC LUPUS ERYTHEMATOSUS PATIENTS FROM ELECTRONIC HEALTH RECORDS Annals of the Rheumatic Diseases 2024;83:1014.