Review

Sjögren's syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts

A.V. Goules[1], T.P. Exarchos[2], V.C. Pezoulas[2], K.D. Kourou[2], A.I. Venetsanopoulou[1], S. De Vita[3], D.I. Fotiadis[2], A.G. Tzioufas[1]

[1]Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens;
[2]Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Greece;
[3]Rheumatology Clinic, DSMB, AOU Santa Maria della Misericordia, University of Udine, Italy.

Andreas V. Goules, MD*
Themis P. Exarchos, PhD*
Vasilis C. Pezoulas, MSc
Konstadina D. Kourou, MSc
Aliki I. Venetsanopoulou, MD
Salvatore De Vita, MD, PhD
Dimitrios I. Fotiadis, PhD
Athanasios G. Tzioufas, MD, PhD

**These authors equally contributed to this paper.*

Please address correspondence to:
Dr Athanasios G. Tzioufas,
Department of Pathophysiology,
School of Medicine,
University of Athens,
Mikras Asias Str 75,
115 27, Athens, Greece.
E-mail: agtzi@med.uoa.gr

Received on November 19, 2018; accepted in revised form on June 18, 2019.

Clin Exp Rheumatol 2019; 37 (Suppl. 118): S175-S184.

Key words: Sjögren's syndrome, harmonisation, biomarkers, unmet needs, data sharing, distributed learning

Funding: the HarmonicsSS project has received funding from the European Union 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.

Competing interests: A.G. Tzioufas is the co-ordinator of the HarmonicSS project. The other co-authors have declared no competing interests.

ABSTRACT

Primary Sjögren's syndrome (pSS) is a chronic, systemic autoimmune disease with diverse clinical picture and outcome. The disease affects primarily middle-aged females and involves the exocrine glands leading to dry mouth and eyes. When the disease extends beyond the exocrine glands (systemic form), certain extraglandular manifestations involving liver, kidney, lungs, peripheral nervous system and the skin may occur. Primary SS is considered the crossroad between autoimmunity and lymphoproliferation, since approximately 5% of patients develop NHL associated lymphomas. As with every chronic disease with complex aetiopathogenesis and clinical heterogeneity, pSS has certain unmet needs that have to be addressed: a) classification and stratification of patients; b) understanding the distinct pathogenetic mechanisms and clinical phenotypes; c) defining and interpreting the real needs of patients regarding the contemporary diagnostic and therapeutic approaches; d) physician and patients' training regarding the wide spectrum of the disease; e) creating common policies across European countries to evaluate and manage SS patients. To achieve these goals, an intense effort is being currently undertaken by the HarmonicSS consortium in order to harmonise and integrate the largest European cohorts of pSS patients. In this review, we present an overview of our perception and vision, as well as new issues arising from this project such as harmonisation protocols and procedures, data sharing principles and various ethical and legal issues originating from these approaches.

Introduction

Primary Sjögren's syndrome (pSS) is a slowly progressive, systemic autoimmune disorder characterised by lymphocytic infiltration around the ductal epithelium of the salivary and lacrimal glands, resulting in mucosal dryness. The syndrome may occur as independent and distinct entity (primary Sjögren's syndrome) or may complicate the course of another systemic autoimmune disease, such as rheumatoid arthritis or lupus (secondary Sjögren's syndrome). Middle-aged women are usually affected, although evidence of the disease may be present years before. The annual incidence of pSS is approximately 7 new cases/100,000 people and the estimated prevalence ranges between 0.01–0.1% of general population, depending upon the cohort nationality, geographical distribution and the applied classification criteria (1-4). The disease has a wide clinical spectrum, extending from a mild and benign gland exocrinopathy confined to the salivary and lacrimal glands to severe and life-threatening conditions, such as vasculitis and lymphoma leading to end-stage organ failure and rarely in death (5, 6). Clinical manifestations of pSS are considered to result from two underlying immunopathologic phenomena: the typical periepithelial lymphocytic infiltration of the affected tissues and the B cell hyperactivity. The involved epithelium of salivary and lacrimal glands has been found to play an important role in disease pathogenesis: a) by expressing intrinsically MHC class I and II, costimulatory and adhesion molecules capable of activating naïve T cells, and b) by producing chemokines and cytokines mediating T and B lymphocyte recruitment and differentiation at the site of lesion. Epithelial cells are also attacked by lymphocytes and die by apoptosis releasing several autoantigens contributing to perpetuation and maintenance of autoimmune local response (7). The central role of the epithelium led to the introduction in the literature of the term "autoimmune epithelitis" as an alternative for pSS (8). The resultant destruction of the epithelium and the replacement by fibrotic structures leads to tissue damage and dysfunction producing mucosal dryness. Similar periepithelial lymphocytic infiltration of the salivary and lacrimal glands can be also observed in other organs such as the bronchi, the kidney and the liver mediating the extraglandular manifestations of the disease (6). B cell hyperactivity, is documented by the presence of hypergammaglobulinaemia and the production of autoantibodies such as antinuclear antibodies (ANA), anti-Ro/SSA, anti-La/SSB and type II cryoglobulins. The ectopic germinal centre (GC)-like structures found within the biopsies of minor salivary glands are thought to represent the active site of cryoglobulins production and probably lymphomagenesis in pSS (9, 10). Type II cryoglobulins associated immune complexes, mediate tissue damage after deposition, producing the so called extraepithelial manifestations such as palpable purpura, peripheral neuropathy and glomerulonephritis (6, 7, 11).

Table I. Classification and prevalence of clinical manifestations in pSS.

Glandular manifestations (%)	Extraglandular manifestations (%)	Extraepithelial manifestations (%)
Dry mouth (>95%)	Periepithelial	Palpable purpura (10%)
Dry eyes (>95%)	Interstitial Nephritis	Glomerulonephritis (2%)
Parotid gland enlargement (~40%)	Incomplete dRTA (~30%)	Peripheral neuropathy (2%)
Vaginal dryness (~30%)	Complete dRTA (5-10%)	
Dry skin (~10%)	Diabetes insipidus (rare)	
	Proximal RTA (rare)	
	Primary Biliary Cholangiitis	
	Bronchitis/bronchiolitis (8%)	
	Non-specific	
	Fatigue (50%)	
	Raynaud's (30%)	
	Arthralgias/arthritis (60%)	
	Interstitial lund disease (2%)	

pSS: primary Sjögren's syndrome.

The clinical spectrum of pSS

In general, pSS clinical manifestations are classified into glandular, extraglandular and extraepithelial immune complex mediated (5, 6) (Table I). Glandular manifestations represent a generalised form of exocrinopathy involving mainly the salivary and lacrimal glands, manifested as mucosal dryness. More than 90% of pSS patients experience oral and eye dryness at the time of diagnosis and almost all of them will develop sicca symptoms during disease course. About 40–50% of pSS patients present with either unilateral or bilateral parotid gland enlargement. In addition, some pSS patients also complain of systemic dryness as part of the generalised exocrinopathy mentioned above, involving the skin, the nose, the trachea as well as the vagina causing dyspareunia. Extraglandular manifestations caused by lymphocytic infiltration around the epithelium of the affected organ include bronchitis/bronchiolitis with persistent irritant dry cough, primary biliary cirrhosis in 3% of pSS patients and interstitial nephritis manifested as incomplete or complete distal renal tubular acidosis (RTA) (6, 12, 13). Although interstitial nephritis is present in 30% of pSS patients in a subclinical form, only 3% among pSS population present with clinically significant renal involvement. Distal RTA leads to alkaline urinary pH, hyposthenuria, hypokalemic hyperchloremic metabolic acidosis with normal anion gap and nephrolithiasis/nephrocalcinosis (6, 12). Proximal RTA and diabetes insipidus are less common clinical manifestations of IN. Other periepithelial extraglandular manifestations include autoimmune thyroid disease in 10–20% of pSS patients leading gradually to hypothyroidism or autoantibody production against thyroid antigens, and autoimmune adrenal and ovarian disease (14). Non-specific extraglandular manifestations are also common among pSS population. Approximately 70% of pSS patients complain of fatigue affecting daily activities and quality of life and 60% develop mild, symmetric non-erosive arthritis/synovitis of small joints. Raynaud's phenomenon, although milder compared to other autoimmune diseases, occurs in 30-40% of patients (14). Almost 10–15% of pSS patients constitute the systemic form of the disease characterised by the appearance of extraepithelial immune complex mediated manifestations and a tendency to develop lymphoma (15, 16). Immune complex mediated vasculitis presenting as palpable purpura of lower extremities or leg ulcers occurs in 10% of patients and is associated with type II cryoglobulinaemia and low C4 complement levels (6). Almost 1-2% of pSS patients present with either mononeuritis multiplex or sensorimotor axonal polyneuropathy in the context of immune complex deposition and subsequent inflammation of the vasa nervosum. Glomerulonephritis with histopathologic features of membranoproliferative or membranous type may also complicate pSS at 2% of patients, producing either nephritic or nephrotic syndrome as a result of cryoglobulins associated immune complex deposition at the glomeruli (12, 15). The life time risk of lymphoma among pSS population is 5% but patients with the systemic form of the disease are more prone to develop lymphoproliferative disorders (17-19). The main lymphoma type is extranodal marginal zone B cell lymphoma of mucosal associated lymphoid tissue (MALT) while diffuse large B cell lymphoma (DLBC) and nodal marginal zone B cell lymphoma are less common types (20).

Prognosis, diagnosis and treatment of pSS

pSS is a benign disease with a good prognosis. For the majority of patients, the clinical picture has been already completed at the time of diagnosis and remains almost unchanged during the disease course (6). Patients with glandular manifestations may experience dryness severe enough to affect their quality of life but without impact on morbidity or mortality. Similarly, extraglandular systemic manifestations evolve slowly, following a benign course with a tendency to occur near disease onset or disease diagnosis (6). In some cases, the chronic and ongoing underlying inflammatory process may lead to slowly progressive organ impairment and end-stage organ failure. On the contrary, extraepithelial immune complex mediated manifestations complicate the disease as a late squeal and these patients display less favourable outcome with increased morbidity and mortality (6, 21). The increased mortality among pSS patients is mainly attributed to the development of lymphoproliferative disorders. Risk factors for lymphoma include persistent or recurrent parotid gland enlargement, palpable purpura, peripheral neuropathy, cryoglobulinaemia, low C4 complement levels and lymphopenia (22-24). The majority of pSS patient seek medical advice many years after the appearance of sicca symptoms (25). However, in some cases the first sign related to the disease may precede mucosal dryness and patients may present with palpable purpura, interstitial nephritis manifested as hyposthenuria, nephrolithiasis/nephrocalcinosis or hypokalemic paralysis, accompanied by the presence of autoantibodies such as anti-Ro/SSA or anti-La/SSB.

The diagnosis of pSS is based on the history, physical examination and evaluation focusing on oral and ocular dryness. Two sets of classification criteria are currently used, although these criteria have been designed for research purposes rather than for routine clinical practice (26, 27). However, many tests incorporated in these sets of criteria, are applied for the diagnostic approach of pSS. A careful history is necessary to identify medical conditions and medications that may cause dryness and mimic pSS such as diabetes mellitus, irradiation of head and neck, bone marrow transplantation leading to graft-*versus*-host disease, viral infections (HIV, HCV, HTLV-1), sarcoidosis, amyloidosis, diuretics and antihistamines for chronic allergic conditions (27). Physical examination may reveal mucosal dryness and in some cases candidiasis of the oral cavity, absence of sublingual salivary pooling, hyperlobulated tongue with loss of papillae and extensive dental carries. Parotid gland or submandibular gland enlargement and generalised lymphadenopathy can be also seen in pSS patients. Other findings from physical examination related to pSS, include palpable purpura of lower extremities and leg ulcers, mononeuritis multiplex or sensorimotor neuropathy as well as signs of nephritic syndrome with edema, dyspnea and hypertension. Specific questionnaires and objective tests are necessary to confirm sicca symptoms (27). For oral dryness, unstimulated whole saliva measurement (positive test defined as: <1.5ml in 15min), sialography or scintigraphy of salivary glands are considered objective tests for hyposalivation and altered structure suggesting hypofunction (27). Other imaging modalities in clinical practice include parotid and submandibular ultrasound and very rarely MRI. Lacrimal dryness can be objectively evaluated by measurement of tear production with the Schirmer's test (positive test defined as: wetting<5mm in 5min) and documentation of keratoconjunctivitis (KCS) either with the Rose Bengal dye and the von Bijsterveld scoring system (≥4/9) (27) or with fluorescein/lissamine green dyes and the SICCA ocular staining score (≥5)(26). Tear break up time is another helpful test to document quality abnormalities of the tear film (normal values >10 sec). Lip (minor labial salivary gland) biopsy is performed to reveal the characteristic histopathologic hallmark of periepithelial infiltration of the salivary epithelium with the presence of at least 1 focus (defined as an aggregate of at least 50 lymphocytes) per 4mm^2 surface area (focus score is defined as the total number of foci per 4mm^2 surface area) (27). Minor labial salivary biopsy is a very useful diagnostic tool to confirm the typical pathologic lesion of pSS and rule out other conditions such as sarcoidosis, IgG4 related disorders, lymphoma, amyloidosis and lipoproteinaemias. Immunologic testing for ANA, ENAs (anti-Ro/SSA, anti-La/SSB, anti-U1RNP, anti-Sm), rheumatoid factors, anti-dsDNA and autoantibodies against thyroid antigens (anti-TG, anti-TPO) should be obtained to confirm the autoimmune nature of the disease and exclude other systemic autoimmune disorders that may accompany pSS such as lupus and mixed connective tissue disease (6, 14). Anti-Ro/SSA antibodies are present in 60-80% of patients while anti-La/SSB in 30–50% and are considered specific for pSS. Other detectable autoantibodies include anti-citrullinated peptides and anti-centromere occurring in small percentage of pSS patients (28, 29). Laboratory tests may show anaemia, leukopenia, thrombocytopenia, hypergammaglobulinaemia, serum monoclonal component, electrolytic and acid-base abnormalities, haematuria, proteinuria and impaired renal function. Therefore, complete blood count with deferential, biochemistry panel, electrolytes, urinalysis, serum protein electrophoresis and chest X ray should be obtained during the initial evaluation of every patient suspected for pSS. Additional investigations for HIV, HCV, HTLV-1 and IgG4 serum levels may be performed in selected cases.

Treatment of pSS is mainly symptomatic aiming to control and alleviate mucosal dryness (30). Patients should be instructed for self-care oral hygiene, scheduled visits to the dentist, good hydration and mechanical salivary flow stimulation (*e.g.* sugar free gums). For severe cases, artificial saliva and muscarinic agonists such as pilocarpine can be prescribed. For ocular dryness, the regular use of artificial tears is highly recommended while moisturisers and lubricants can be used to alleviate vaginal dryness and dyspareunia. For the systemic manifestations of the disease, an organ based management is recommended. Hydroxychloroquine and methotrexate can be administered for arthralgias/arthritis while calcium channel blockers are the first choice for controlling Raynaud's phenomenon (14). Fatigue can be managed with SSRI's and regular exercise (14). Inhaled β2 agonists are prescribed for patients with bronchitis or bronchiolitis and oral alkali potassium bicarbonate supplements, depending on the type of interstitial nephritis, can be given accordingly (14, 15). Prednisone, azathioprine or iv cyclophosphamide are reserved for cases of interstitial lung disease. Immune complex mediated extraepithelial manifestations are treated with prednisone or steroid pulses, azathioprine, anti-B cell depletion therapies, IV cyclophosphamide, IVIG and plasmapheresis, depending on the severity and the type of organ involvement (14, 15). Sjögren's syndrome associated lymphomas display a favourable prognosis with the majority of patients having high 5-year survival rates (20). Depending on the histologic type, the stage and the individualised characteristics of each patient (*e.g.* age, comorbidities, etc.), management include watch and wait policy for the majority of patients with MALT lymphomas or rituximab with or without an alkylating agent for more severe cases while the R-CHOP regimen is the first treatment choice for pSS patients with DLBC lymphomas (20).

Table II. The unmet needs for pSS.

- Common European policies for evaluation and management of pSS patients
- Design common diagnostic and therapeutic approaches for pSS patients
- Train physicians and patients about the disease spectrum
- Study and understand critical pathogenetic mechanisms of the disease
- Novel biomarkers
  - Early diagnosis and patients' stratification
  - Patients' classification according to disease phenotypes and endotypes
  - Identification of novel therapeutic targets
  - Prognostic and therapeutic response markers
  - Molecular predictors of lymphoma

pSS: primary Sjögren's syndrome.

Unmet needs of pSS

In the past two decades, the revolutionary explosion of biotechnology and the progress in bioinformatics, allowed the SS scientific community to achieve very critical steps in terms of understanding and familiarising with the clinical needs of SS. One of the first important steps was the consensus on the classification criteria that took place on 2002 (27). This proposed classification, allowed the various clinical and basic investigators to recruit patients for research purposes based on well-defined subjective and objective parameters of the disease, ensuring relative homogeneity among the various cohorts so that the conclusions could be generalised. Recently, another set of classification criteria was constructed taking into account also the systemic nature of pSS (26). Extensive clinical studies conducted by highly experienced researchers managed to describe the natural history of the disease with its various outcomes and to classify the clinical manifestations into distinct subgroups, defining in this way the main clinical phenotypes of the syndrome. It became clear soon, that patients with the systemic form of the disease exhibit increased mortality and morbidity and this particular subset gained further clinical attention, mainly because of the high risk of lymphoma. Following and analysing a growing body of data, several clinical and laboratory indices were identified as lymphoma and mortality predictors in an effort to therapeutically manage and modify the adverse outcomes. In parallel, based on the experience from other autoimmune diseases, modern biologic treatments were tested in large cohorts of pSS patients to control the disease and improve patients' symptoms, unfortunately without much success, although some improvement was observed in certain parameters of pSS. Criticisms on the methodology of these studies revealed that specific features of pSS were ignored: a) the slowly progressive disease course that leads to already advanced stages at the time of diagnosis, as a result of chronic and well established pathogenetic mechanisms; b) the short observation time of these studies as opposed to the slowly evolving nature of pSS; c) the heterogeneity regarding the phenotypes and endotypes of recruited patients; d) the lack of in-depth understanding of the underlying cellular, molecular and intracellular mechanisms that drive pathogenesis. However, new efficacy data from ongoing pSS studies with promising biologic agents are still pending.

The relative inefficacy of biologic agents in pSS raised concerns about the orientation and the goals of research and pointed out a transition period towards an era of precision medicine according to reliable biomarkers. The unmet needs at the clinical and research level are defined by the prospectives of biotechnology on the one hand, and the specific features of pSS on the other hand (Table II). The scientific community should therefore aim at the following goals: a) to reveal the important pathogenetic mechanisms that are implicated in the initiation, establishment, maintenance, damage and repair of the affected tissues; b) to correlate the distinct clinical phenotypes with the underlying cellular and molecular events that predominate and define each subset of patients; c) to identify reliable biomarkers.

The discovery of novel biomarkers is of great clinical importance. A deeper understanding of the cellular and molecular mechanisms will allow us to identify critical molecules that could potentially serve as biomarkers of the disease. Ideally, the biomarker associated variables, should be easily measured at the proper biological specimen (*e.g.* saliva, serum, homogenised tissue specimen from minor salivary glands) after performing a simple, cheap and reproducible methodology (31).

Given that the candidate biomarkers really reflect important biological aspects of the disease, are expected to contribute to: a) early diagnosis and identification of pSS patients at the initial stages of the disease; b) sub-classification of patients according to clinical phenotypes and endotypes of the disease; c) identification of possible therapeutic targets and estimation of response to treatment; d) prediction of lymphoma development and efficient follow-up; e) more sophisticated stratification and recruitment of patient for future clinical trials.

The initial approach for identifying novel biomarkers should be based on preliminary clinical stratification of pSS patients and subsequent investigation of key molecules that could potentially serve to characterise the various clinical subsets of the disease. A complementary approach, exploring genetic or other molecular biomarkers by direct application on the whole pSS population, should be also considered but for future use. To meet all the aforementioned unmet needs for pSS and proceed smoothly in the era of precision medicine, it is very important to develop the largest possible cohort of patients reflecting the underlying national, geographical and environmental heterogeneity as well as the phenotypic and endotypic diversity of the disease. A large harmonised cohort is expected to: a) facilitate the identification of reliable biomarkers through extensive bioinformatics analysis and subsequent validation studies; b) create tools to be used for the stratification of patients; c) offer training skills to physicians; d) develop common health policies for the disease. In order to achieve these goals, the scientific community needs large integrated cohorts that should be harmonised in a way to provide the maximal information for analytic purposes.

The HarmonicSS project, funded by the European Commission, gathers the most important clinical and research centres of pSS in Europe along with highly experienced engineers and bioinformaticians, to successfully harmonise the maximum possible data from all the involved clinical partners and offer new tools for large data analysis.

Data sharing and data governance

To permit federated and distributed access to various cohorts for analysis, several obstacles must be overcome. A data provider (*e.g.* a clinical centre) wishing to share clinical data to a healthcare platform must provide all the necessary ethical and legal documents, prior to any further processing, including: (i) legitimate interests; (ii) data protection impact assessment; (iii) the purpose of processing; (iv) signed consents to the processing of personal data from the data subjects; (v) purpose of transferring to third parties; (vi) data protection guarantee; (vii) notifications to the data subject about the processing, among many others. To obtain access to multiple cohorts, nowadays, a cloud platform can be introduced.

In general, a cloud platform should take into consideration several technical challenges. Standardisation for defining a common format for the clinical datasets, categorisation for terminology description, and harmonisation through ontology alignment are only but a few technical limitations. Data clearing (referred to as data curation) for deduplication (*i.e.* the identification of duplicate variables), data imputation (*i.e.* automated methods to fill missing values), outlier detection is important prior to data harmonisation. Other technical challenges include: (i) informed consent forms for data sharing through handshaking (Fig. 1) and pooled data analysis (handshaking), (ii) encrypting algorithms for secure execution of all the data analytics services within the cloud, (iii) ethical issues for data collection introduced by different countries inside and outside the EU, (iv) data protection (*e.g.* according to the GDPR guidelines in Europe or the HIPAA guidelines in USA), (v) fear for data abuse, (vi) data monitoring, validation, and storage, (vii) multidimensional interoperability (legal, regulatory, applications, IT infrastructures), (viii) the cost scalability over security which is a crucial trade-off, and (ix) effective data curation mechanisms (*e.g.* for dealing with duplicate fields, missing values, outliers within the clinical data). In addition, the definition of the primary data collectors (*i.e.* data providers) and secondary analysts (*i.e.* data processors) must be clarified.

The heterogeneity of the clinical data across different countries (*e.g.* structural and vocabulary dissimilarities) can be overcome by de-centralised analysis. Data aggregation (*i.e.* data pseudonymisation using hashing or tokenisation) is mandatory prior to data sharing. Software and Creative Commons (CC) licenses must be obtained for all the tools that have been implemented within the cloud. The definition of a common data model that is able to describe the domain knowledge of a disease (*e.g.* parameters and relations between them, vocabularies) is also necessary for semantic matching during data harmonisation. The complexity of the hospital IT infrastructures poses significant barriers towards the de-centralised processing of the data. Finally, the existence of a committee which orchestrates the data acquisition and data processing (cloud shepherds) is necessary for providing guidelines towards the realisation of a federated platform.

Medical data sharing involves the above mechanisms concerning the protection of patient's rights and privacy. In addition, the data governance framework is responsible for the evaluation of the quality and completeness of the clinical data by taking into consideration existing public health policies. The data sharing mechanism along with the data governance framework constitute the two fundamental mechanisms prior to the development and application of the data harmonisation and analytics services.

De-centralised data harmonisation

Data harmonisation is a data-driven approach which aims to overcome the heterogeneity of the cohorts worldwide by converting the heterogeneous datasets into compatible ones, with minimum loss. Data harmonisation involves several mechanisms including dataset description, dataset transformation, similarity description, terminology detection and alignment, semantic and lexical matching, etc. In addition, the main idea behind medical data harmonisation is based on the mapping of each template of interest into a (predefined) common reference template. According to the literature, there are two types of matching heterogeneous datasets, namely the lexical matching and the semantic matching (30). Here, emphasis will be given only on semantic matching since it combines lexical matching (*i.e.* string matching) along with the associations between the variables of the heterogeneous datasets. Semantic matching is usually conducted by mapping the ontology of a dataset (source ontology) into the ontology of the other's (target ontology) based on semantic interlinking approaches. In order to ensure a patient data protection compliant harmonisation ap-

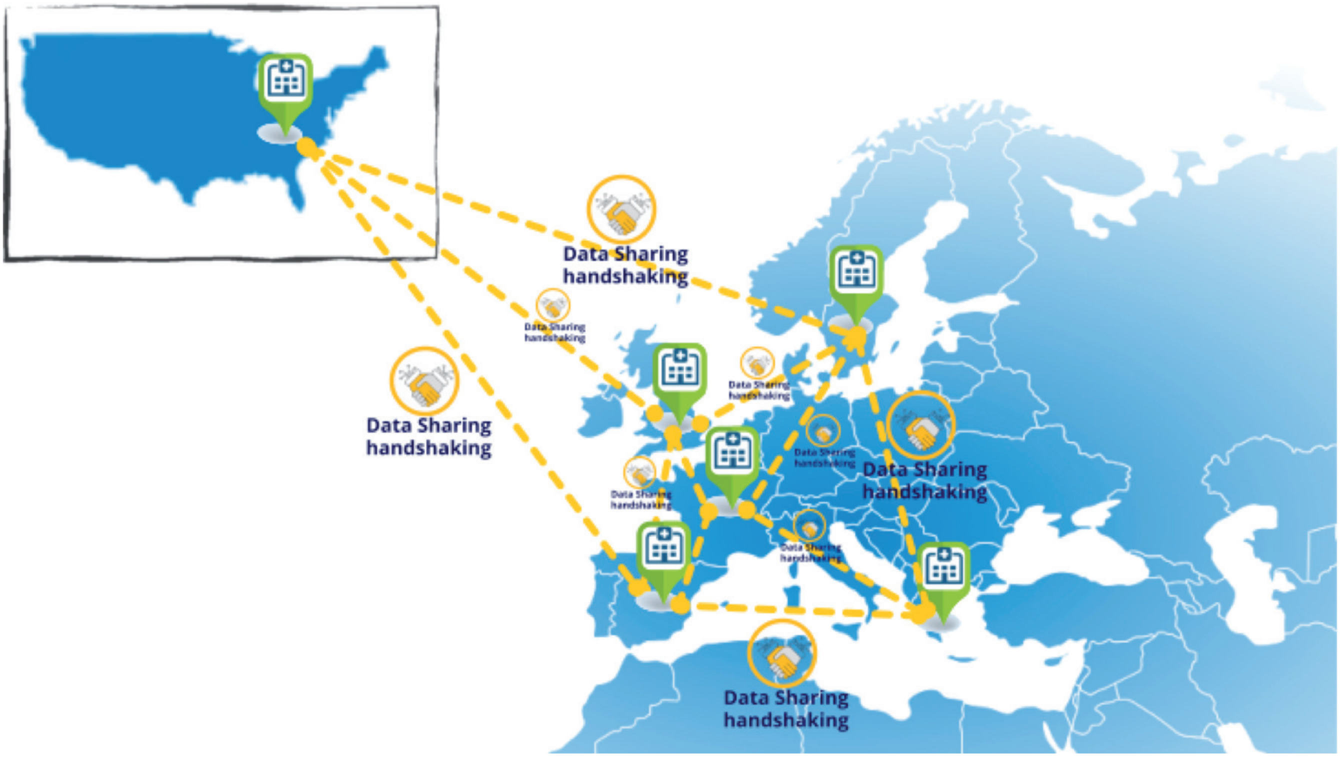

Fig. 1. Illustration of the distributed analysis strategy for data sharing - handshaking. Each involved cohort must fulfill all the necessary data protection requirements for data sharing. In the case of cross-border data sharing, additional legal requirements should be taken into consideration (*e.g.* between the General Data Protection Regulation guidelines in Europe and the Health Insurance Portability and Accountability Act guidelines in USA). In addition, prior to any analysis of the cohort data, appropriate requests for data access are sent from the data analysts to the data providers that manage the cohorts of interest. The data providers can either accept or reject these requests. This procedure aims to reassure the privacy of the data and is referred to as handshaking. The data analytics services can gain access to the private space of each individual cohort only when the handshaking is succesfull.

proach, the data harmonisation shall be performed on each cohort's private space. The harmonised data are then stored on these private spaces for further analysis. Examples of tools which are often employed for harmonisation are presented in the sequel. Interestingly, all these softwares and technical tools are open access and not commercially available.

The Opal software (32) can be used for data harmonisation based on a predefined template (reference model) in order to enable data integration across the cohorts. The Mica software (32) is a client-server application developed to (i) construct web portals for individual studies, (ii) create a study catalogue or registry based on the input data, and (iii) enable data access and management. In addition, the SORTA tool (33) is another software tool for data re-coding and data annotation as well. The DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) software tool (34) is a dynamically evolving entity that combines tools from OBiBa (*i.e.* Mica for developing web portals for studies and Opal for data harmonisation and integration). It is mainly comprised by two fundamental platforms, namely the DataSchema Platform and the Harmonization Platform. The concept of (de-centralised) data harmonisation involves the following mechanisms: (i) cohort data description, (ii) cohort data transformation, (iii) ontology alignment, and (iv) the existence of a knowledge database.

Initially, cohort data description refers to the complete description of all the variables within the cohort's dataset (*e.g.* the terminologies and type of each variable). Cohort data transformation is the process where each dataset is transformed to a common format for better manipulation. Then, each transformed dataset is converted to an ontology which serves as the source ontology for the ontology alignment procedure. An ontology describes the domain knowledge of a disease and is usually expressed in .OWL (Web Ontology Language) or .RDFS (Resource Description Framework Schema) format. Protégé is a widely used tool for constructing ontologies and frameworks for health systems (35). In fact, each source ontology is mapped to a target ontology. The target ontology is the ontology which is constructed from the reference model and comprises the core of the de-centralised data harmonisation procedure. The reference model is a complete description of the minimum requirements (domain knowledge) that a complete dataset should fulfill and is provided by the clinical experts of the disease's domain.

Ontology alignment is also known as semantic interlinking or ontology mapping. An ontology mapping algorithm uses a source ontology and a target ontology as input and produces an aligned version of the source ontology based on the target ontology, *i.e.* a set of possible matches for each variable of the source ontology with those from the target ontology. This procedure is semi-automatic since the clinician's assistance

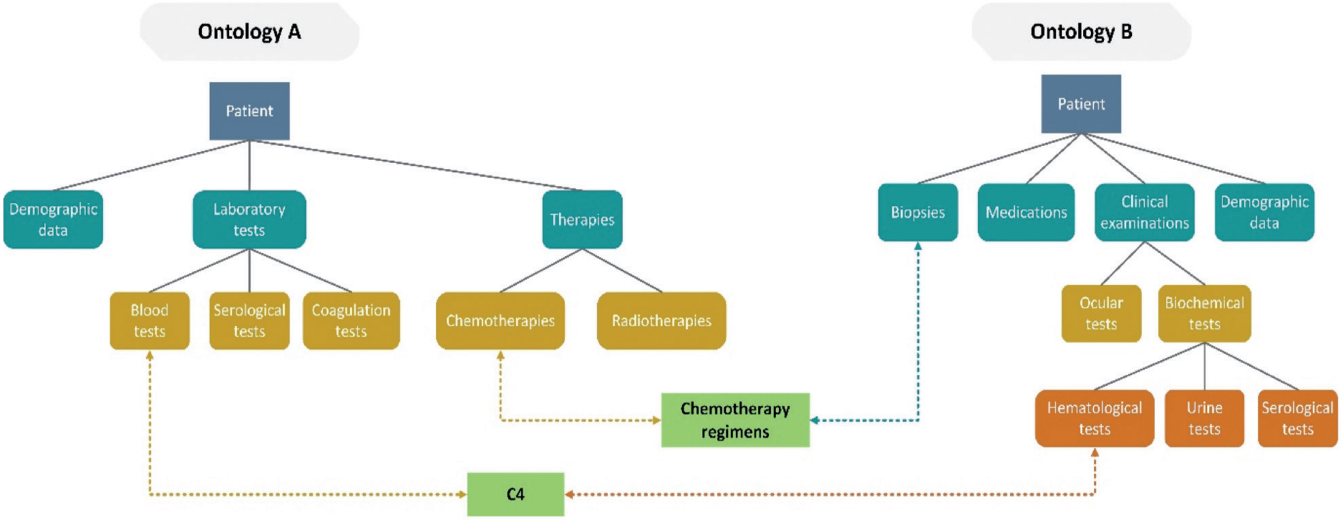

Fig. 2. A simple illustration of the semantic matching procedure. The main objective of the semantic matching algorithm is to seek for associations between the terms of two ontologies A and B that are instances of larger ontologies which describe the domain knowledge of the Sjögren's syndrome. In this example, the algorithm seeks for relational matches between these two ontologies for the terms 'C4' and 'Chemotherapy regimens'. In Ontology A, the term 'C4' is defined as a parameter in the subclass 'Blood tests' of the class 'Laboratory tests' which is connected with the main class 'Patient' through an object property. In Ontology B, however, the term 'C4' is located in the subclass 'Haematological tests' of the class 'Biochemical tests' which in turn is a subclass of the class 'Clinical examinations' that is further connected with the main class 'Patient'. The algorithm uses the semantic relations (object properties) and the related vocabularies (blood, haematological) to find a relational match between these two subclasses and yeild a relational association for the term 'C4'. In a similar way, the algorithm uses the semantic relations and the related vocabularies (chemotherapies, biopsies) to identify a relational match between the classes 'Biopsies' and 'Chemotherapies' which yields a relational association for the term 'Chemotherapy regimens'.

is necessary in order to select the most appropriate terms for each variable. The ontology mapping file can be expressed in EDOAL (Expressive Declarative Ontology Alignment Language), HTML (Hypertext Markup Language) formats. Existing ontology matching tools include the S-Match (36). The knowledge database is a collection of existing ontologies and vocabularies which are related to the domain of the disease under examination and reduces the clinician's involvement during the ontology alignment procedure. More specifically, this database can be combined with the upcoming ontologies in order to train a "smart" system that will be able to execute automatic ontology mapping using machine learning.

A simple example of the semantic matching process is presented in Figure 2. Ontologies A and B are instances of larger ontologies which describe the domain knowledge of SS from two different cohorts. The semantic matching algorithm uses structural and vocabulary information in order to identify a relational match (*i.e.* an association) between the subclass 'Blood tests' (which belongs to the broader class named 'Laboratory Tests') in Ontology A and the subclass 'Haematological tests' in Ontology B, which yields a relational association for the term 'C4'. In a similar manner, the algorithm uses again relational knowledge to relationally match the subclass 'Chemotherapies' (which belongs to the broader class named 'Therapies') in Ontology A and the subclass 'Biopsies' in Ontology B which yields a relational association for the term 'Chemotherapy regimens'.

Distributed data analytics from federated databases

The field of data analytics consists of a variety of services including tools for extracting knowledge (*i.e.* mining knowledge) from medical big data. The data mining algorithms consist of supervised and unsupervised machine learning algorithms which are widely employed for constructing prediction models (*e.g.* a lymphoma prediction model), as well as, testing these models using various performance evaluation measures (*e.g.* accuracy, ROC curves). The concept of distributed data analytics is based on the fact that the data do not leave the space they reside. According to this concept, the initial data mining algorithm is distributed from a reference node to all the other nodes for training and testing purposes (Fig. 3A). These computing nodes can be viewed as the private spaces of the clinical centres. On each node, the algorithm is locally executed (Fig. 3B) and the results return to the reference node (Fig. 3C). The results are finally combined to the reference node and distributed to all the involved nodes (Fig. 3D). Through this way, the analysis is secure but biases are often introduced into the results due to: (i) the heterogeneous structure of the data, and (ii) the type of the algorithm (*e.g.* linear or non-linear). The first factor can be overcome by harmonising the involved data. As far as the second factor is concerned, emphasis must be given to the development of appropriately-defined distributed algorithms which is an ongoing research field of interest. At this point, it is important to note that whenever a researcher wishes to apply data models (or algorithms) on the clinical data, he/she shall first request (and grant) permission from the corresponding data providers of these cohorts.

A straightforward infrastructure that meets the distributed analysis requirements is presented in Figure 1 (37). This infrastructure provides a general framework for executing distributed learning algorithms by taking into con-

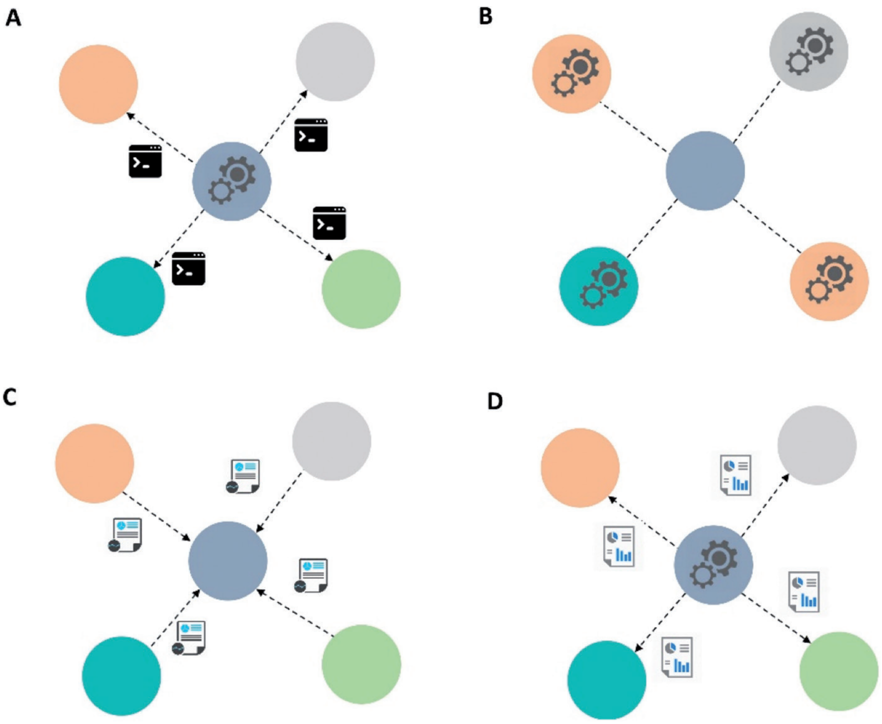

Fig. 3. Steps towards privacy preserving distributed learning. The private spaces of the clinical centres can be viewed as computing nodes which are located in multiple sites (*e.g.* on a cloud) and receive commands from a central node. The commands may vary from the computation of simple descriptive statistics (*e.g.* histograms) to the construction of data models for prediction (*e.g.* lymhpoma prediction). Here, we assume that the command is the execution of a properly selected data mining algorithm for prediction. (**A**) The data mining algorithm is distributed from the central node to the rest of the nodes. (**B**) The data mining algorithm is executed locally (trained and tested) on the data of each node, in a parallel manner. (**C**) The performance evaluation results from each node return to the central node. (**D**) The individual results are finally aggregated (combined) on the central node and the final global prediction model is distributed to the rest of the nodes for further evaluation.

sideration all the necessary data sharing issues (including the ethical and legal ones). Several distributed learning schemes have been proposed including the soft-margin *l*1-regularised space Support Vector Machine (38) or combined Bayesian network models (39). Conventional cross-validation approaches can also be used for performance evaluation (*e.g.* k-fold cross validation) so as to extract various performance indicators. The results of these studies are very promising; a fact that enhances the effectiveness of the distributed analysis concept.

Data protection

One of the major issues to consider when sharing and harmonising data from different cohorts is the protection of ethical and legal aspects of the shared and harmonised data. Towards this several different aspects and methodologies/policies have been identified by either the clinical centres participating in data sharing or expert legal offices that have been assigned the role of solving the legal constraints to share data between different countries with different data protection laws and regulations. The innate sensitivity of biomedical information has led to a set of principles and rules for safeguarding protection and privacy of personal data in all stages of data manipulation starting from data collection protocols to data analysis infrastructures.

The evolution of big data and their multi-purpose utilisation towards new knowledge mining, which is enabled by innovative developments in information technologies (IT), pose new challenges that need to be addressed in the context of informed consent, privacy of data, ownership, and epistemology in assessing big data ethics and objectivity of big data (40). For instance, a comprehensive essay centreing on issues regarding consent obtaining in biobank studies recognised the need for a defensible, sustainable and conceptually coherent consent policy (10). Adapting freely given, specific and informed consent along with anonymisation mechanisms such that: (i) facilitating the process of dynamic re-consent though the use of IT providing a transparency level between individuals and their data and, (ii) balancing the need for irreversible anonymisation and data linkage and continuing data update, are key issues in the protection of individuals with regard to processing of personal data (40-45).

European Union General Data Protection Regulation (GDPR)

The intensified interest on big data sharing, aggregation, linkage and analysis yielded to the forthcoming replacement of the European Union (EU) Data Protection Directive (DPD) 95/46/EC by the General Data Protection Regulation (GDPR). At the heart of the EU DPD and GDPR lie the principles of fairness and lawfulness assuring the openness and legality of the use of personal sensitive data. The GDPR (http://www.eugdpr.org/eugdpr.org.html) mainly overhauls the EU Directive 95/46/EC with respect to rights of the data subject by introducing or strengthening the rights to: (i) access to data, (ii) rectification and erasure ('right to be forgotten'), (iii) data portability, and (iv) notification for a personal data breach. In addition, the concept of privacy by design calls for the effective implementation of appropriate technical and organisational measures (*e.g.* pseudo-anonymisation) from the design phase of a system in order to ensure non-attribution to an identified or identifiable natural person and meet the requirements of GDPR. The conditions for consent have been also strengthened requesting clarity of the information provided to the individuals. According to the GDPR, every medical-based architecture should take into consideration data protection by design and by default principles. More specifically, there are four roles that should be clearly described; the data provider (or data controller), the data processor, the data owner (*i.e.* patient) and the Data Protection Officer (DPO).

A data provider who wishes to distribute data should provide all necessary signed consent forms, data protection guarantees, recipients of personal data, legitimate interests concerning the data subject, the purpose of processing, the contact details of the DPO, etc. A patient has the right to be forgotten and the right for data portability.

Framework for Responsible Sharing of Genomic and Health-Related Data

The EU BioSHARE Project has developed, under the aegis of the Global Alliance for Genomics and Health, the Framework for Responsible Sharing of Genomic and Health-Related Data (46-49). This Framework has established a set of foundational principles for responsible sharing of genomic and health-related data: (i) respect individuals, families and communities, (ii) advance research and scientific knowledge, (iii) promote health, wellbeing and the fair distribution of benefits; and (iv) foster trust, integrity and reciprocity. In addition, it has set out ten core elements complementing the interpretation of the aforementioned principles: (i) transparency, (ii) accountability, (iii) engagement, (iv) data quality and security, (v) privacy, data protection and confidentiality, (vi) risk-benefit analysis, (vii) recognition and attribution, (viii) sustainability, (ix) education and training, and (x) accessibility and dissemination.

Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act (HIPAA) of 1996, Public Law 104-191, which is part of the Social Security Act, aims to embrace the sharing of certain patient administrative data for promoting the healthcare industry (50). HIPAA is comprised by two fundamental Rules, namely the Privacy Rule and the Security Rule. The latter sets national standards for the protection of personal health information that is created, received, used, or maintained by a covered entity (50). The Privacy Rule establishes national standards for the protection of healthcare electronic transactions including medical records and other personal health information that are conducted by healthcare providers. The covered entities must ensure the: (i) confidentiality, (ii) integrity, and (iii) availability of the digital health information they create, receive, maintain or transmit and identify and protect against any threats to the security or integrity of the information.

References

1. GORANSSON LG, HALDORSEN K, BRUN JG *et al.*: The point prevalence of clinically relevant primary Sjögren's syndrome in two Norwegian counties. *Scand J Rheumatol* 2011; 40: 221-4.
2. MALDINI C, SEROR R, FAIN O *et al.*: Epidemiology of primary Sjögren's syndrome in a French multiracial/multiethnic area. *Arthritis Care Res* 2014; 66: 454-63.
3. QIN B, WANG J, YANG Z *et al.*: Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Ann Rheum Dis* 2015; 74: 1983-9.
4. THOMAS E, HAY EM, HAJEER A, SILMAN AJ: Sjögren's syndrome: a community-based study of prevalence and impact. *Br J Rheumatol* 1998; 37: 1069-76.
5. GOULES AV, TZIOUFAS AG: Primary Sjögren's syndrome: Clinical phenotypes, outcome and the development of biomarkers. *Autoimmun Rev* 2016; 15: 695-703.
6. SKOPOULI FN, DAFNI U, IOANNIDIS JP, MOUTSOPOULOS HM: Clinical evolution, and morbidity and mortality of primary Sjögren's syndrome. *Semin Arthritis Rheum* 2000; 29: 296-304.
7. GOULES AV, KAPSOGEORGOU EK, TZIOUFAS AG: Insight into pathogenesis of Sjögren's syndrome: Dissection on autoimmune infiltrates and epithelial cells. *Clin Immunol* 2017; 182: 30-40.
8. MOUTSOPOULOS HM: Sjögren's syndrome: autoimmune epithelitis. *Clin Immunol Immunopathol* 1994; 72: 162-5.
9. THEANDER E, VASAITIS L, BAECKLUND E *et al.*: Lymphoid organisation in labial salivary gland biopsies is a possible predictor for the development of malignant lymphoma in primary Sjögren's syndrome. *Ann Rheum Dis* 2011; 70: 1363-8.
10. KROESE FGM, HAACKE EA, BOMBARDIERI M: The role of salivary gland histopathology in primary Sjögren's syndrome: promises and pitfalls. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S222-33.
11. TZIOUFAS AG, MANOUSSAKIS MN, COSTELLO R, SILIS M, PAPADOPOULOS NM, MOUTSOPOULOS HM: Cryoglobulinemia in autoimmune rheumatic diseases. Evidence of circulating monoclonal cryoglobulins in patients with primary Sjögren's syndrome. *Arthritis Rheum* 1986; 29: 1098-104.
12. GOULES A, MASOURIDI S, TZIOUFAS AG, IOANNIDIS JP, SKOPOULI FN, MOUTSOPOULOS HM: Clinically significant and biopsy-documented renal involvement in primary Sjögren syndrome. *Medicine* (Baltimore) 2000; 79: 241-9.
13. KAMPOLIS CF, FRAGKIOUDAKI S, MAVRAGANI CP, ZORMPALA A, SAMAKOVLI A, MOUTSOPOULOS HM: Prevalence and spectrum of symptomatic pulmonary involvement in primary Sjögren's syndrome. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S94-101.
14. MAVRAGANI CP, MOUTSOPOULOS HM: Sjögren syndrome. *CMAJ* 2014; 186: E579-86.
15. GOULES AV, TATOULI IP, MOUTSOPOULOS HM, TZIOUFAS AG: Clinically significant renal involvement in primary Sjögren's syndrome: clinical presentation and outcome. *Arthritis Rheum* 2013; 65: 2945-53.
16. PAPAGEORGIOU A, VOULGARELIS M, TZIOUFAS AG: Clinical picture, outcome and predictive factors of lymphoma in Sjögren syndrome. *Autoimmun Rev* 2015; 14: 641-9.
17. VOULGARELIS M, DAFNI UG, ISENBERG DA, MOUTSOPOULOS HM: Malignant lymphoma in primary Sjögren's syndrome: a multicenter, retrospective, clinical study by the European Concerted Action on Sjögren's Syndrome. *Arthritis Rheum* 1999; 42: 1765-72.
18. VOULGARELIS M, ZIAKAS PD, PAPAGEORGIOU A, BAIMPA E, TZIOUFAS AG, MOUTSOPOULOS HM: Prognosis and outcome of non-Hodgkin lymphoma in primary Sjögren syndrome. *Medicine* (Baltimore) 2012; 91: 1-9.
19. ZINTZARAS E, VOULGARELIS M, MOUTSOPOULOS HM: The risk of lymphoma development in autoimmune diseases: a meta-analysis. *Arch Intern Med* 2005; 165: 2337-44.
20. PAPAGEORGIOU A, ZIOGAS DC, MAVRAGANI CP *et al.*: Predicting the outcome of Sjögren's syndrome-associated non-hodgkin's lymphoma patients. *PLoS One* 2015; 10: e0116189.
21. IOANNIDIS JP, VASSILIOU VA, MOUTSOPOULOS HM: Long-term risk of mortality and lymphoproliferative disease and predictive classification of primary Sjögren's syndrome. *Arthritis Rheum* 2002; 46: 741-7.
22. BRITO-ZERON P, RAMOS-CASALS M, BOVE A, SENTIS J, FONT J: Predicting adverse outcomes in primary Sjögren's syndrome: identification of prognostic factors. *Rheumatology* (Oxford) 2007; 46: 1359-62.
23. NISHISHINYA MB, PEREDA CA, MUNOZ-FERNANDEZ S *et al.*: Identification of lymphoma predictors in patients with primary Sjögren's syndrome: a systematic literature review and meta-analysis. *Rheumatol Int* 2015; 35: 17-26.
24. TZIOUFAS AG, BOUMBA DS, SKOPOULI FN, MOUTSOPOULOS HM: Mixed monoclonal cryoglobulinemia and monoclonal rheumatoid factor cross-reactive idiotypes as predictive factors for the development of lymphoma in primary Sjögren's syndrome. *Arthritis Rheum* 1996; 39: 767-72.
25. PAVLIDIS NA, KARSH J, MOUTSOPOULOS HM: The clinical picture of primary Sjögren's syndrome: a retrospective study. *J Rheumatol* 1982; 9: 685-90.
26. SHIBOSKI CH, SHIBOSKI SC, SEROR R *et al.*: 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: A consensus and data-driven methodology involving three international patient

cohorts. *Ann Rheum Dis* 2017; 76: 9-16.

27. VITALI C, BOMBARDIERI S, JONSSON R *et al.*: Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann Rheum Dis* 2002; 61: 554-8.
28. BAER AN, MEDRANO L, MCADAMS-DEMARCO M, GNIADEK TJ: Association of anticentromere antibodies with more severe exocrine glandular dysfunction in Sjögren's syndrome: analysis of the sjögren's international collaborative clinical alliance cohort. *Arthritis Care Res* (Hoboken) 2016; 68: 1554-9.
29. RYU YS, PARK SH, LEE J *et al.*: Follow-up of primary Sjögren's syndrome patients presenting positive anti-cyclic citrullinated peptides antibody. *Rheumatol Int* 2013; 33: 1443-6.
30. MAVRAGANI CP, NEZOS A, MOUTSOPOULOS HM: New advances in the classification, pathogenesis and treatment of Sjögren's syndrome. *Curr Opin Rheumatol* 2013; 25: 623-29.
31. ARGYROPOULOU OD, VALENTINI E, FERRO F *et al.*: One year in review 2018: Sjögren's syndrome. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S14-26.
32. DOIRON D, BURTON P, MARCON Y *et al.*: Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol* 2013; 10: 12.
33. PANG C, SOLLIE A, SIJTSMA A *et al.*: SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. *Database* (Oxford) 2015; 2015.
34. FORTIER I, BURTON PR, ROBSON PJ *et al.*: Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010; 39: 1383-93.
35. UNIVESITY S: Protege: a free, open source ontology editor and knowledge-base framework.
36. GIUNCHIGLIA F, AUTAYEU A, PANE J: S-Match: an open source framework for matching lightweight ontologies. *Semantic Web* 2012; 3: 307-17.
37. DEIST TM, JOCHEMS A, VAN SOEST J *et al.*: Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017; 4: 24-31.
38. BRISIMI TS, CHEN R, MELA T, OLSHEVSKY A, PASCHALIDIS IC, SHI W: Federated learning of predictive models from federated Electronic Health Records. *Int J Med Inform* 2018; 112: 59-67.
39. JOCHEMS A, DEIST TM, EL NAQA I *et al.*: Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. *Int J Radiat Oncol Biol Phys* 2017; 99: 344-52.
40. MITTELSTADT BD, FLORIDI L: The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics* 2016; 22: 303-41.
41. BULL S, ROBERTS N, PARKER M: Views of ethical best practices in sharing individual-level data from medical and public health research: A systematic scoping review. *J Empir Res Hum Res Ethics* 2015; 10: 225-38.
42. CAPOCASA M, ANAGNOSTOU P, D'ABRAMO F *et al.*: Samples and data accessibility in research biobanks: an explorative survey. *Peer J* 2016; 4: e1613.
43. CAULFIELD T, MURDOCH B: Genes, cells, and biobanks: Yes, there's still a consent problem. *PLoS Biol* 2017 Jul; 15: e2002654.
44. HALLINAN D, FRIEDEWALD M: Open consent, biobanking and data protection law: can open consent be 'informed'under the forthcoming data protection regulation? *Life Sci Soc Policy* 2015; 11: 1.
45. MOSTERT M, BREDENOORD AL, BIESAART MC, VAN DELDEN JJ: Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach. *Eur J Hum Genet* 2016; 24: 956.
46. KNOPPERS BM: International ethics harmonization and the global alliance for genomics and health. *Genome Med* 2014; 6: 13.
47. KNOPPERS BM: Framework for responsible sharing of genomic and health-related data. *HUGO J* 2014; 8: 3.
48. KNOPPERS BM, HARRIS JR, BUDIN-LJØSNE I, DOVE ES: A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum Genet* 2014; 133: 895-903.
49. RAHIMZADEH V, DYKE SO, KNOPPERS BM: An international framework for data sharing: Moving forward with the global alliance for genomics and health. *Biopreserv Biobank* 2016; 14: 256-9.
50. ATCHINSON BK, FOX DM: From the field: the politics of the Health Insurance Portability And Accountability Act. *Health Aff* (Milwood) 1997; 16: 146-50.