Editorial

The Big Data Sjögren Consortium: a project for a new data science era

N. Acar-Denizli[1], B. Kostov[2,3], M. Ramos-Casals[4-6], on behalf of the Sjögren Big Data Consortium

[1]Department of Statistics, Faculty of Science and Letters, Mimar Sinan Güzel Sanatlar Üniversitesi, Istanbul, Turkey; [2]Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain; [3]Primary Healthcare Transversal Research Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; [4]Sjögren Syndrome Research Group (AGAUR), Laboratory of Autoimmune Diseases Josep Font, IDIBAPS-CELLEX, Barcelona, Spain; [5]Department of Autoimmune Diseases, ICMiD, Hospital Clínic, Barcelona, Spain; [6]Department of Medicine, University of Barcelona, Spain.

Nihan Acar-Denizli, PhD
Belchin Kostov, PhD
Manuel Ramos-Casals, MD, PhD

Please address correspondence to:
Dr Nihan Acar-Denizli,
Department of Statistics,
Faculty of Science and Letters,
Mimar Sinan Güzel Sanatlar Üniversitesi,
Cumhuriyet Mh. Silahşör Cd. 89,
Bomonti, 34380 Şişli/İstanbul, Turkey.
E-mail: nihan.acar@msgsu.edu.tr

Received on May 6, 2019; accepted in revised form on July 1, 2019.

Clin Exp Rheumatol 2019; 37 (Suppl. 118): S19-S23.

Key words: primary Sjögren's syndrome, big data, data science, data sharing, data visualisation

The members of the EULAR-SS Task Force Big Data Consortium are listed in the Appendix.

Competing interests: none declared.

In the 21st century, data science has taken a new meaning, thanks to big data, and almost all the people who are dedicating to investigate will have to learn how to deal with this. Volume (amount of data), variety (number of types of data) and velocity (speed of data processing) are considered as three main defining properties of big data, called 3 Vs of Big Data, according to many experts in Data Science (1). Extended definitions include veracity (quality or trustworthiness of the data) and value (the worth of the data being extracted) as the fourth and fifth V, respectively. However, one of them stands out from the rest: volume. When someone mentions "big data", we immediately think about a huge volume of data, with the definition of the term "huge" being dependent on the specific research area. In Big Data era, the use of innovative and appropriate techniques to deal with this huge amount of data has been emerged substituting classical statistical techniques. Nowadays, almost everyone talks about machine learning, data mining or artificial intelligence, even there is a widespread misuse of these terms. For example, data mining is widely used to refer big data despite the fact that two terms correspond to different concepts. The classical definition of Data Mining refers to extracting knowledge from data. Although the term is frequently used in studies including Big Data analysis, data mining techniques may also be used to extract knowledge from smaller data sets. Likewise, artificial intelligence could be basically defined as the study of creating intelligent agents and machine learning as the science of creating algorithms and programmes which learn on their own. In fact, we all use in our day-to-day life machine learning algorithms in one way or the other and probably we don't even realise it. Virtual personal assistants, traffic predictions, online transportation networks, social media services, online customer support and product recommendations are just a few examples. There are a plenty of applications of machine learning algorithms and data mining techniques in marketing, business, finance, security, engineering, and many other fields where biomedical research is not an exemption. Medicine was identified early as one of the most promising application areas for artificial intelligence where researchers have proposed and developed many clinical decision support systems that have been shown to interpret electrocardiograms, diagnose diseases, choose appropriate treatments, provide interpretations of clinical reasoning and assist physicians in generating diagnostic hypotheses in complex patient cases (2). In this sense, both clinical and biomedical researchers, including physicians, nurse practitioners, clinical pharmacist, physician assistants, biologists and chemists, are changing the way they did research up to now. Sharing clinical data across hospitals to support open innovation is an old idea, but which is being taken up by the scientific community at an increasing speed, concerning public sharing in particular (3). The uncertainty in data protection in general, and with respect to international transfer in particular, could generate some doubts around the ethic limitations of clinical data sharing. In this sense, the General Data Protection Regulation (GDPR), recently approved by the European Parliament, introduces an exemption to the general prohibition for the processing of sensitive personal data. According to the 'research exemption' in Ar-

ticle 9(2)(j) under which sensitive personal data, including genetic data, can be processed without adhering to the strict consent requirements (4). In other words, whatever purpose the sensitive personal data was collected for, further processing for research purposes without the data subject's consent is considered compatible with the initial purposes. However, this exemption should be taken with caution as processing sensitive data under the research exemption requires appropriate safeguards such as data minimisation, anonymisation, and data security. Moreover, the data subject possesses several rights such as the right of access by the data subject, the right to rectification, the right to restriction of processing or the right to object in order to derogate personal data processing under the research exemption.

Big data is transforming the way of doing clinical and biomedical research (5). Nowadays, national and international collaborations that allow creating big data registries are substituting studies carried out only in one centre with limited sample sizes. The Big Data Sjögren Project Consortium was born out of this idea: this international, multicentre registry was designed in 2014 to take a "high-definition" picture of the main features of primary Sjögren's syndrome (SS) using worldwide data-sharing cooperative merging of pre-existing clinical SS databases from leading centres in clinical research in SS from the five continents. The centres share a harmonised data infrastructure and conduct cooperative online efforts in order to refine already-collected data in each centre. The codebook containing instructions on the variables and data codification was firstly discussed and approved by the Steering Committee members, and was further shared with the consortium partners. Databases from each centre were harmonised into a single database by applying specific pre-processing techniques such as the detection and treatment of outliers, influential observations, errors and missing data. By January 2019, the participant centres had included 11,421 valid patients from 24 countries (Fig. 1). This international scientific collaboration, designed according to a

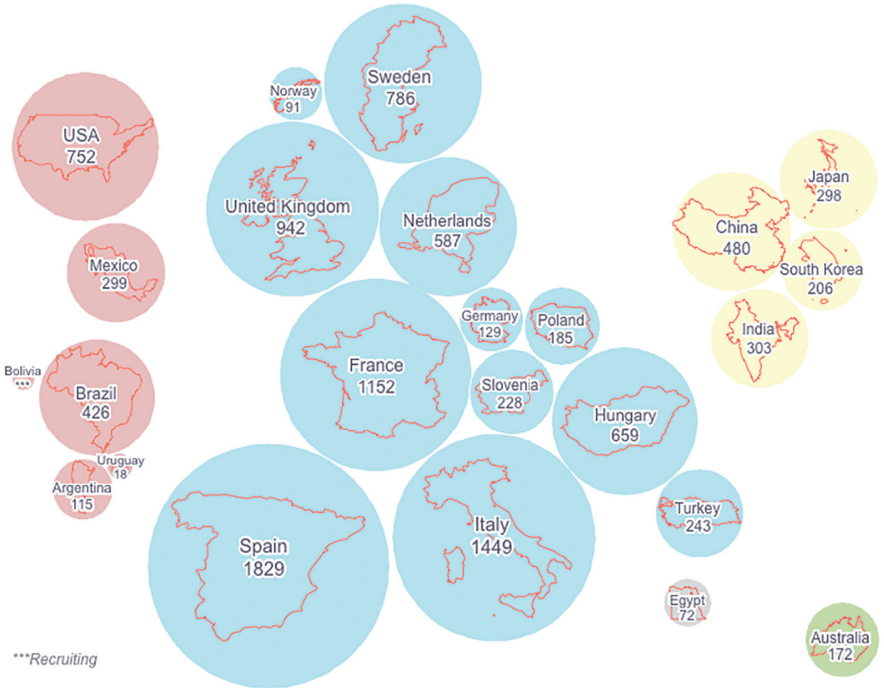

Fig. 1. The Big Data Sjögren Project Consortium in January 2019: 45 cohorts from 24 countries of the 5 continents; 11,421 patients included fulfilling the 2002 criteria for primary SS.

'Data Sharing' approach, has produced excellent results until now in spite of several limitations associated to retrospective design, predominant presence of European patients, the magnitude of the selection bias between the different countries, different medical practices across regions, and assays used by the different centres. Despite these limitations, the results obtained can be used to improve the clinical management of patients with primary SS. The Big Data Sjögren Project provided the first evidence for a strong influence of geolocation and ethnicity on the phenotype of primary SS at diagnosis (6). It also confirmed a strong influence of immunological markers on the phenotype of primary SS at diagnosis in the largest multi-ethnic international cohort ever analysed, with a greater influence for cryoglobulinaemic-related markers in comparison with Ro/La autoantibodies and ANA (7).

While the availability of big data offers many possibilities for enhancing our knowledge about human diseases, the need for a cautious use of statistical concepts is essential and raises many challenges (8). The fact that as clinical researchers we have to deal with big data sets brought the focus of attention on one of the most frequently discussed controversies in medical research: interpretation of the results based mainly on statistical significance. As we all know very well, studies with big data sets may detect some differences which, although statistically significant, may not be relevant clinically. This debate, which is not new, has gained strength lately. A recently published special issue "Statistical inference in the 21st century: a world beyond $p<0.05$" by The American Statistician, an official journal of the American Statistical Association, with more than 40 original papers discusses topics such as getting to a post "$p<0.05$" era, interpreting and using p, supplementing or replacing p and changing publication policies and statistical education (9). All this clearly highlights that it is time that researchers should start to look for patterns, trends, and associations and, to stop focusing only on p-values and statistical significances that are generally superfluous in order to analyse big data sets. So, what can researchers do? The answer is easy: give much more attention to the clinical interpretation of the results and to prioritise and extend the use of data visualisation techniques for presenting the results, which are actu-

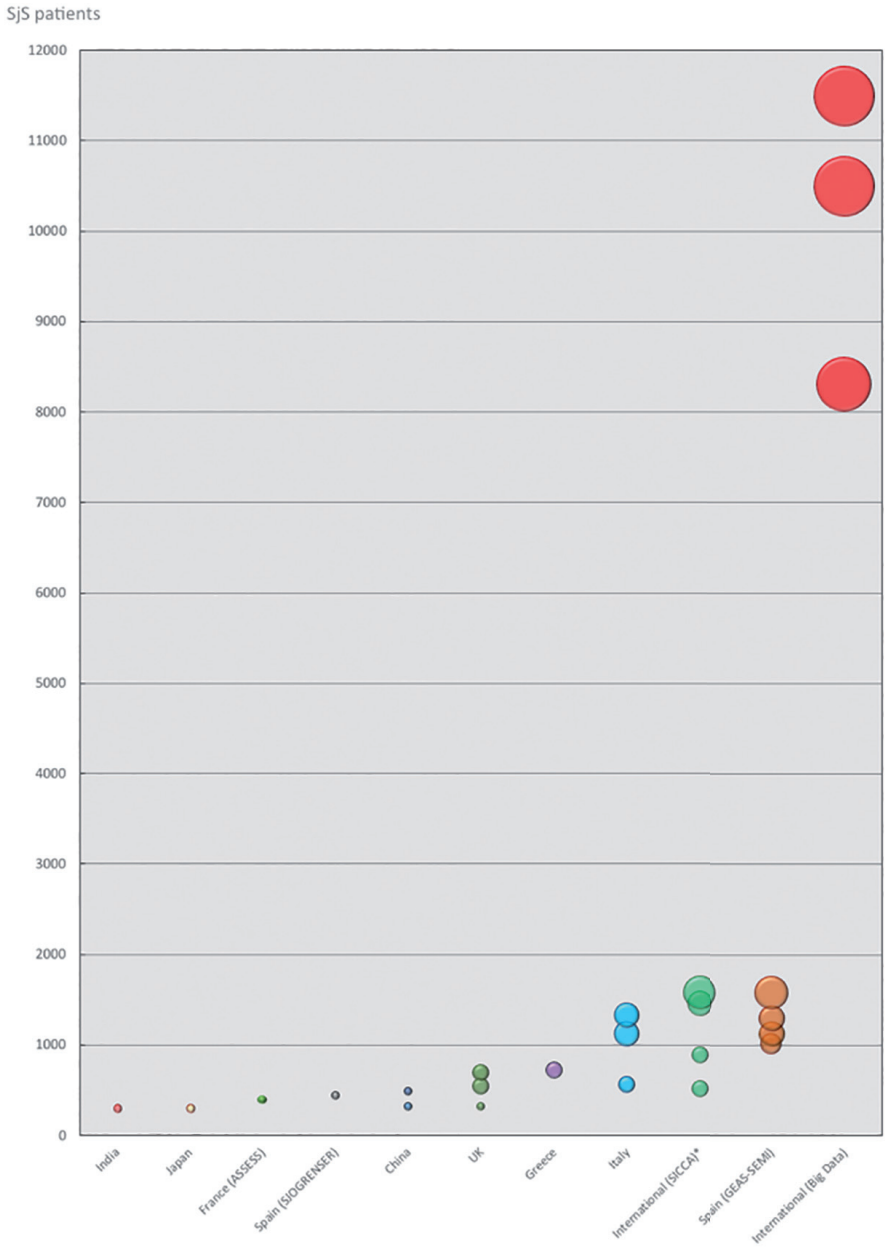

Fig. 2. Number of patients with primary SS included in international registries and multicentric cohorts. *Also including associated SS.

ally one of the pivotal tools supporting the results obtained by Big Data analysis. In the actual world, to represent analyses of millions of patients and thousands of variables, classical results tables are not a solution. Instead, scientists should use novel data visualisation techniques such as bubble plots, heatmaps, spider charts, wordclouds, network diagrams, choropleth maps, etc. that allow representing information from Big Data sets. As has been stated by Elgendi (10), data visualisation skills greatly affect research quality and the publications and there are many examples in the current literature that indicate either a lack of knowledge or lack of a concerted effort toward the proper use of data visualisation techniques. Most scientific journals, especially those with a high impact factor, are now requiring visualisations that allow a greater impact of the results published in papers. This means that researchers have to devote a huge effort to find the most appropriate graphs for their data such as an original manuscript published in The *New England Journal of Medicine* that includes four figures and no table (11).

The Sjögren Big Data Project has managed to multiply by 6 the largest number of patients ever included in an international registry, and by 10 the largest number of patients included in nationwide multicentric cohorts (Fig. 2). The project is currently analysing the clinical and geoepidemiological characterisation of the disease in patients from the 5 continents, including people of different ethnicities, something that had not been achieved previously. In addition to the evident breakthrough in the knowledge of the worldwide impact of the disease, the project has contributed to achieve other goals that are equally important for the scientific community devoted to the study of Sjögren. All this has been possible thanks to the creation of a multicentre registry that today includes nearly 12,000 patients for a disease with a prevalence lower than 0.1%. The individual contribution of every participating centre, even those that included a modest number of cases, continues to increase progressively year by year, which reflects that participation in the project serves as an enhancer of the clinical experience and acknowledgement that each centre acquires at a regional/national level. And the almost 5 years of close collaboration has created a solid nucleus of clinical research, led by internists and rheumatologists, but also with the active participation of many other medical specialties, with more than 100 people currently involved, including the active participation in each centre of young researchers, who will be the future experts on SS. This is a group that has also achieved bidirectional communication between the data scientists (people who have the knowledge of maths and statistics and programming skills to conduct sophisticated and systematic analyses of data) and the clinicians, an essential feature to reach reliable clinical conclusions from Big Data studies. This close collaboration between researchers and data scientists is the key factor to achieve success in big data projects. Thus, the team of mathematicians and statisticians have progressively integrated a clinical vision in their data management design while, at

the same time, clinicians have learned to use in their hypotheses and methodologies the new concepts of big data, data mining and the new statistics and visualisation techniques, thus changing the way of thinking and analysing, and moving on beyond the search for the classical (and probably obsolete) *p*-value difference – a cooperative model of clinical and epidemiological research, sustained and validated by data scientists, that can be easily replicable in other systemic autoimmune diseases, whose complexity and heterogeneity could also be successfully addressed through the use of Big Data Sharing.

APPENDIX

Members of the EULAR-SS Task Force Big Data Consortium

a) *Members of the EULAR-SS Task Force*

P. Brito-Zerón, C. Morcillo (Autoimmune Diseases Unit, Department of Medicine, Hospital CIMA- Sanitas, Barcelona, Spain); P. Brito-Zerón, A. Flores-Chávez, M. Ramos-Casals (Sjögren Syndrome Research Group (óAGAUR), Laboratory of Autoimmune Diseases Josep Font, IDIBAPS-CELLEX, Department of Autoimmune Diseases, ICMiD, University of Barcelona, Hospital Clínic, Barcelona, Spain); N. Acar-Denizli (Department of Statistics, Faculty of Science and Letters, Mimar Sinan Fine Arts University, Istanbul, Turkey); F. Ng (Institute of Cellular Medicine, Newcastle University, Newcastle Upon Tyne, UK); Ildike-Fanny Horvath, Antónia Szántó (Division of Clinical Immunology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary); A. Rasmussen, K. Sivils, H. Scofield (Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA); R. Seror, X. Mariette (Center for Immunology of Viral Infections and Autoimmune Diseases, Assistance Publique – Hôpitaux de Paris, Hôpitaux Universitaires Paris-Sud, Le Kremlin-Bicêtre, Université Paris Sud, INSERM, Paris, France Paris, France); T. Mandl, P. Olsson (Department of Rheumatology, Malmö University Hospital, Lund University, Lund, Sweden); X. Li, B. Xu (Department of Rheumatology and Immunology, Anhui Provincial Hospital, China); C. Baldini, S. Bombardieri (Rheumatology Unit, University of Pisa, Pisa, Italy); J.E. Gottenberg (Department of Rheumatology, Strasbourg University Hospital, Université de Strasbourg, CNRS, Strasbourg, France); D. Danda, P. Sandhya (Department of Clinical Immunology & Rheumatology, Christian Medical College & Hospital, Vellore, India); L. Quartuccio, L. Corazza, S De Vita (Clinic of Rheumatology, Department of Medical and Biological Sciences, University Hospital "Santa Maria della Misericordia", Udine, Italy); R. Priori, A. Minniti (Department of Internal Medicine and Medical Specialties, Rheumatology Clinic, Sapienza University of Rome, Italy); G. Hernandez-Molina, J. Sánchez-Guerrero (Immunology and Rheumatology Department, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán. México City, Mexico); A.A. Kruize, E. van der Heijden (Department of Rheumatology and Clinical Immunology, University Medical Center Utrecht, Utrecht, The Netherlands); V. Valim (Department of Medicine, Federal University of Espírito Santo, Vitória, Brazil); M. Kvarnstrom, M. Wahren-Herlenius (Department of Medicine, Solna, Division of Experimental Rheumatology, Karolinska Institutet, and Karolinska University Hospital, Stockholm); D. Sene (Service de Médecine Interne 2, Hôpital Lariboisière, Université Paris VII, Assistance Publique-Hôpitaux de Paris, 2, Paris, France); R. Gerli, E. Bartoloni (Rheumatology Unit, Department of Medicine, University of Perugia, Italy); S. Praprotnik (Department of Rheumatology, University Medical Centre, Ljubljana, Slovenia); D. Isenberg (Centre for Rheumatology, Division of Medicine , University College London , UK); R. Solans (Department of Internal Medicine, Hospital Vall d'Hebron, Barcelona, Spain); M. Rischmueller, S. Downie-Doyle (Department of Rheumatology, The Queen Elizabeth Hospital and University of Adelaide, South Australia, Australia); S-K. Kwok, S-H. Park (Division of Rheumatology, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea); G. Nordmark (Rheumatology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden); Y. Suzuki, M. Kawano (Division of Rheumatology, Kanazawa University Hospital , Kanazawa , Ishikawa , Japan); R. Giacomelli, F. Carubbi (Clinical Unit of Rheumatology, University of l'Aquila, School of Medicine, L'Aquila, Italy); V. Devauchelle-Pensec, A. Saraux (Rheumatology Department, Brest University Hospital, Brest, France); M. Bombardieri, E. Astorri (Centre for Experimental Medicine and Rheumatology, Queen Mary University of London, UK); B. Hofauer, A. Knopf (Otorhinolaryngology / Head and Neck Surgery, Technical University Munich, Munich, Germany); H. Bootsma, A. Vissink (Department of Rheumatology & Clinical Immunology, University of Groningen, University Medical Center Groningen, the Netherlands); D. Hammenfors, J.G. Brun (Department of Rheumatology, Haukeland University Hospital, Bergen, Norway); G. Fraile (Department of Internal Medicine, Hospital Ramón y Cajal, Madrid, Spain); S. E. Carsons (Division of Rheumatology, Allergy and Immunology Winthrop-University Hospital, Stony Brook University School of Medicine, Mineola, NY, USA); T. A. Gheita, (Rheumatology Department, Kasr Al Ainy School of Medicine, Cairo University, Egypt); E.M. Abd El-Latif (Ophthalmology Department, Faculty of Medicine, Alexandria University, Egypt); H.M. Khalil (Ophthalmology Department, Faculty of Medicine, Beni Suef University, Egypt); J. Morel (Department of Rheumatology, Teaching hospital and University of Montpellier, Montpellier, France); C. Vollenveider (German Hospital, Buenos Aires, Argentina); F. Atzeni (IRCCS Galeazzi Orthopedic Institute, Milan and Rheumatology Unit, University of Messina, Messina, Italy); S. Retamozo (Instituto De Investigaciones En Ciencias De La Salud (INICSA), Universidad Nacional de Córdoba (UNC), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) - Córdoba - Argentina. Instituto Universitario de Ciencias Biomédicas de Córdoba (IUCBC), Córdoba- Argentina); V. Moça Trevisano (Federal University of São Paulo, Sao Paulo, Brazil); B. Armagan, L. Kilic, U. Kalyoncu (Department of Internal Medicine, Hacettepe University, Faculty of Medicine, Ankara, Turkey); H. Nakamura, T. Shimizu, A. Takatani (Department of Immunology and Rheumatology, Division of Advanced Preventive Medical Sciences, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan); T. Nakamura, Y. Takagi (Department of Radiology and Cancer Biology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan); A. Sebastian, P. Wiland (Department of Rheumatology and Internal Medicine, Wroclaw Medical Hospital, Wroclaw, Poland); S.G. Pasoto (Rheumatology Division, Hospital das Clinicas, Faculdade de Medicina da Universidade de Sao Paulo (HCFMUSP), Sao Paulo, Brazil); B. Kostov, A. Sisó-Almirall (Primary Healthcare Transversal Research Group, IDIBAPS, Centre d'Assistència Primària ABS Les Corts, CAPSBE, Barcelona, Spain); S. Consani-Fernández (Internal Medicine, Hospital Maciel, Montevideo, Uruguay. Universidad de la República (UdelaR), Montevideo, Uruguay).

b) *Members of the French ASSESS Cohort*
J. Sibilia (Rheumatology Centre National de Référence des Maladies Auto-Immunes Rares, Institut National de la Santé et de la Recherche Médicale UMRS_1109, Fédération de Médecine Translationnelle de Strasbourg, Strasbourg University Hospital, Université de Strasbourg, Strasbourg, France); C. Miceli-Richard, G. Nocturne (Rheumatology, Bicetre Hospital, Institut National de la Santé et de la Recherche Médicale U-1012, Université Paris Sud, Assistance Publique des Hôpitaux de Paris, Paris, France); J. Benessiano (Centre de Ressources Biologiques, Bichat Hospital, Assistance Publique des Hôpitaux de Paris, Paris, France); P. Dieude (Rheumatology, Bichat Hospital, Assistance Publique des Hôpitaux de Paris, Paris, France); J-J. Dubost (Rheumatology, Clermont-Ferrand Hospital, Clermont-Ferrand, France); A-L. Fauchais (Internal Medicine, Limoges Hospital, Limoges, France); V. Goeb (Rheumatology, Amiens University Hospital, Amiens, France); E. Hachulla (Pierre Yves Hatron, Internal Medicine, Lille University Hospital, Lille, France); C. Larroche (Internal Medicine, Avicenne Hospital, Assistance Publique des Hôpitaux de Paris, Bobigny, France); V. Le Guern, X. Puéchal (Internal Medicine, Cochin Hospital, Assistance Publique des Hôpitaux de Paris, Paris, France); J. Morel (Rheumatology, Montpellier University Hospital, Montpellier, France);
A. Perdriger (Rheumatology, Rennes University Hospital, Rennes, France); S. Rist, Rheumatology, Orléans Hospital, Orléans, France; O. Vittecoq (Rheumatology, Rouen University Hospital, Rouen, France); P. Ravaud (Centre of Clinical Epidemiology, Hotel Dieu Hospital, Assistance Publique des Hôpitaux de Paris, Institut National de la Santé et de la Recherche Médicale U378, University of Paris Descartes, Faculty of Medicine, Paris, France).

c) *Members of the Spanish GEAS Cohort (SS Study Group, Autoimmune Diseases Study Group GEAS, Spanish Society of Internal Medicine SEMI)*
B. Díaz-López (Department of Internal Medicine, Hospital Universitario Central de Asturias, Oviedo), C. Feijoo, (Department of Internal Medicine, Hospital Parc Taulí, Sabadell), L. Pallarés (Department of Internal Medicine, Hospital Son Espases, Palma de Mallorca), M. López-Dupla (Department of Internal Medicine, Hospital Joan XXIII, Tarragona), R. Pérez-Alvarez (Department of Internal Medicine, Hospital do Meixoeiro, Vigo), M. Ripoll (Department of Internal Medicine, Hospital Infanta Sofía, Madrid), B. Pinilla (Department of Internal Medicine, Hospital Gregorio Marañón, Madrid), M. Akasbi (Department of Internal Medicine, Hospital Infanta Leonor, Madrid), B. Maure (Department of Internal Medicine, Complejo Hospitalario Universitario, Vigo), E. Fonseca (Department of Internal Medicine, Hospital de Cabueñes, Gijón), J. Canora (Department of Internal Medicine, Hospital Universitario de Fuenlabrada, Madrid), G de la Red (Department of Internal Medicine, Hospital Espíritu Santo, Barcelona), A.J. Chamorro (Department of Internal Medicine, Complejo Hospitalario de Ourense, Ourense), I. Jiménez-Heredia (Department of Internal Medicine, Hospital de Manises, Valencia, Spain), P. Fanlo (Complejo Universitario de Navarra), P. Guisado-Vasco (Hospital Quirón, Madrid), M. Zamora (Hospital Virgen de las Nieves, Granada).

References

1. MAYER-SCHÖNBERGER V, KENNETH C: Big data: A revolution that will transform how we live, work, and think. *Houghton Mifflin Harcourt* 2013.
2. YU KH, BEAM AL, KOHANE IS: Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; 2: 719-31.
3. CONRADO DJ, KARLSSON MO, ROMERO K, SARR C, WILKINS JJ: Open innovation: Towards sharing of data, models and workflows. *Eur J Pharm Sci* 2017; 109S: S65-71.
4. PORMEISTER K: Genetic data and the research exemption: is the GDPR going too far?. *International Data Privacy Law* 2017; 7: 137-46.
5. OBERMEYER Z, EMANUEL EJ: Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016; 375: 1216-9.
6. BRITO-ZERON P, ACAR-DENIZLI N, ZEHER M *et al*. Influence of geolocation and ethnicity on the phenotypic expression of primary Sjögren's syndrome at diagnosis in 8310 patients: a cross-sectional study from the Big Data Sjögren Project Consortium. *Ann Rheum Dis* 2017; 76: 1042-50.
7. BRITO-ZERON P, ACAR-DENIZLI N, NG WF *et al*.: How immunological profile drives clinical phenotype of primary Sjögren's syndrome at diagnosis: analysis of 10,500 patients (Sjögren Big Data Project). *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S102-12.
8. COX DR, KARTSONAKI C, KEOGH RH: Big data: Some statistical issues. *Stat Probab Lett* 2018; 136: 111-5.
9. WASSERSTEIN RL, SCHIRM AL, LAZAR NA: Moving to a world beyond "p<0.05". *Am Stat* 2017; 73 (Suppl. 1): 1-19.
10. ELGENDI M: Scientists need data visualization training. *Nat Biotechnol* 2017; 35: 990-91.
11. AFSHIN A, FOROUZANFAR MH, REITSMA MB *et al*.; GBD 2015 OBESITY COLLABORATORS: Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med* 2017; 377: 13-27.