Autoimmunity Reviews 14 (2015) 670–679

Contents lists available at ScienceDirect

Autoimmunity Reviews

journal homepage: www.elsevier.com/locate/autrev

Review

# Google-driven search for big data in autoimmune geoepidemiology: Analysis of 394,827 patients with systemic autoimmune diseases

Manuel Ramos-Casals [a,*], Pilar Brito-Zerón [a,1], Belchin Kostov [b], Antoni Sisó-Almirall [b], Xavier Bosch [c], David Buss [a], Antoni Trilla [d], John H. Stone [e,f], Munther A. Khamashta [g], Yehuda Shoenfeld [h]

[a] Josep Font Laboratory of Autoimmune Diseases, CELLEX, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Department of Autoimmune Diseases, ICMiD, Hospital Clínic, Barcelona, Spain
[b] Primary Care Research Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Primary Care Centre Les Corts, CAPSE, Barcelona, Spain
[c] Department of Internal Medicine, ICMiD, Hospital Clínic, Barcelona, Spain
[d] Preventive Medicine and Epidemiology Unit, Hospital Clínic-Universitat de Barcelona, Barcelona Centre for International Health Research, Barcelona, Catalonia, Spain
[e] Harvard Medical School, Boston, MA 02114, USA
[f] Department of Medicine, Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital, Boston, MA 02114, USA
[g] Lupus Research Unit, The Rayne Institute, St Thomas' Hospital, King's College University, London, UK
[h] Zabludowicz Center for Autoimmune Diseases, Chaim Sheba Medical Center, Tel Hashomer, Israel Incumbent of the Laura Schwarz-Kipp Chair for Research of Autoimmune Diseases, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

*Article history:*
Received 16 March 2015
Accepted 30 March 2015
Available online 2 April 2015

*Keywords:*
Geoepidemiology
Big data
Google
Systemic lupus erythematosus
Sjögren syndrome
Vasculitis
Systemic sclerosis

SUMMARY

Systemic autoimmune diseases (SADs) are a significant cause of morbidity and mortality worldwide, although their epidemiological profile varies significantly country by country. We explored the potential of the Google search engine to collect and merge large series (>1000 patients) of SADs reported in the Pubmed library, with the aim of obtaining a high-definition geoepidemiological picture of each disease. We collected data from 394,827 patients with SADs. Analysis showed a predominance of medical vs. administrative databases (74% vs. 26%), public health system vs. health insurance resources (88% vs. 12%) and patient-based vs. population-based designs (82% vs. 18%). The most unbalanced gender ratio was found in primary Sjögren syndrome (pSS), with nearly 10 females affected per 1 male, followed by systemic lupus erythematosus (SLE), systemic sclerosis (SSc) and antiphospholipid syndrome (APS) (ratio of nearly 5:1). Each disease predominantly affects a specific age group: children (Kawasaki disease, primary immunodeficiencies and Schonlein–Henoch disease), young people (SLE Behçet disease and sarcoidosis), middle-aged people (SSc, vasculitis and pSS) and the elderly (amyloidosis, polymyalgia rheumatica, and giant cell arteritis). We found significant differences in the geographical distribution of studies for each disease, and a higher frequency of the three SADs with available data (SLE, inflammatory myopathies and Kawasaki disease) in African-American patients. Using a "big data" approach enabled hitherto unseen connections in SADs to emerge.



**Contents**



* Corresponding author at: Servei de Malalties Autoimmunes, Hospital Clínic, C/Villarroel, 170, 08036 Barcelona, Spain. Tel.: +34 93 2275774; fax: +34 93 2271707.
*E-mail address:* mramos@clinic.ub.es (M. Ramos-Casals).
[1] Supported by the Josep Font Research Fellow Award, Hospital Clinic, Barcelona, Spain.

http://dx.doi.org/10.1016/j.autrev.2015.03.008
1568-9972/

## 1. Introduction

The term "big data" was first used in 1997 in a paper written by Michael Cox and David Ellsworth from NASA, describing the problem they had with visualization (i.e. computer graphics) of large data sets [1]. Ten years later, in a book entitled "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society", American computer scientists popularized the term, predicting that big-data computing will transform the activities of companies, scientific researchers and medical practitioners [2]. More recently, the term has been redefined by Tom Davenport ("the broad range of new and massive data types that have appeared over the last decade or so") and by Gill Press ("the new tools helping us find relevant data and analyze its implications") [3,4].

The use of big data in medicine has received a growing and enthusiastic support in recent years. Medical care is rapidly changing due to the major technical advancements developed for the capture, collection and analysis of large medical databases like electronic health records (EHRs) and personal health records (PHRs). The potential of big data-driven medical research is huge, especially in areas such as epidemiology, genetics and therapeutics, and ranges from relatively basic analysis (such as determining disease incidences in a given population) to more complex studies (efficacy of new drugs in a real world setting, large-scale genetics research). For prevalent diseases, the major advantage of big data is that the larger the study, the more likely that the research findings are closer to the real population. However, for infrequent/rare diseases, in which the inherent patient-by-patient clinical variability often leads to a heterogeneous description of the disease between reported studies, big data-driven research may help to obtain a disease picture with a more "high-definition resolution".

## 2. Autoimmune geoepidemiology

Autoimmune diseases are a significant cause of morbidity and mortality and might affect 5–10% of the world population [5]. Autoimmune diseases are classified as organ-specific (the autoimmune damage is overwhelmingly focused on a specific organ) or systemic (SADs, when the disease affects a large number of organs and systems). Epidemiologically, most SADs are classified as rare diseases (frequency < 5 cases per 10,000 people) except for Sjögren's syndrome and systemic lupus erythematosus. SADs are characterized by a wide spectrum of demographic patterns with respect to the age at diagnosis, gender distribution and ethnic differences. Studying the distribution of these diseases across various geographic regions and ethnic groups may help advance our understanding of the corresponding genetic and environmental underpinnings [6].

## 3. The use of Google search

This review explored the potential of the Google search engine to collect and merge large series (>1000 patients) of SADs reported in the Pubmed library, with the aim of obtaining a high-definition geoepidemiological picture of each disease. We made a text-word search in Google (www.google.com) between the 20th and 31st of January 2015 using the following text algorithm: "systemic autoimmune disease" and "1000…100,000,000 patients" and "www.ncbi.nlm.nih.gov". Table 1 summarizes the list of systemic autoimmune diseases that were individually searched for. Inclusion criteria were defined as the description of at least 1000 patients with a well-defined SAD. Reviews, meta-analysis, epidemiological studies in non-SAD patients and duplicated cohorts were excluded. Two reviewers (MRC and PBC) independently examined the entries retrieved by the Google search for potential eligibility. Studies marked for possible inclusion by either reviewer underwent dual, independent full-text review in Pubmed. If reviewers disagreed, conflicts were resolved by consensus. A data extraction form was developed prior to manuscript review to gather relevant geoepidemiological data from each article. Variables collected were systemic autoimmune disease, number of patients, country, data sources, type of database, inclusion and exclusion criteria, gender distribution, mean age at onset, diagnosis, and/or study entry, and ethnicity data. No restrictions were placed on language or type of publication. We manually searched reference lists of selected articles for relevant citations missed by the search. Possible overlapping data was managed by contrasting the following variables between studies: name of authors, participant centers, number of patients, epidemiological features, type of organ involvement, period of patient recruitment, and

**Table 1**
List of systemic autoimmune diseases (SADs) analyzed: number of studies and patients per disease.

| SADs[a] | Studies (n) | Patients (n) | Patients excluded (n) | Patients analyzed (n) |
|---|---|---|---|---|
| Systemic lupus erythematosus | 13 | 93,070 | 19,091 | 74,081 |
| Kawasaki disease | 7 | 76,119 | 0 | 76,119 |
| Giant cell arteritis | 6 | 59,171 | 4758 | 54,413 |
| Behçet disease | 12 | 37,583 | 2823 | 34,760 |
| Systemic sclerosis | 12 | 32,067 | 482 | 31,585 |
| Primary Sjögren syndrome | 7 | 24,997 | 3301 | 21,696 |
| Sarcoidosis | 9 | 24,433 | 3157 | 21,276 |
| Primary immunodeficiencies | 5 | 20,830 | 0 | 20,830 |
| Amyloidosis | 6 | 11,058 | 0 | 11,058 |
| Medium-sized vasculitis | 2 | 4665 | 692 | 3973 |
| Polymyalgia rheumatica | 1 | 3925 | 676 | 3249 |
| Inflammatory myopathies | 1 | 2477 | 0 | 2477 |
| Granulomatosis with polyangiitis | 2 | 2200 | 111 | 2089 |
| Schonlein–Henoch disease | 1 | 1232 | 0 | 1232 |
| Antiphospholipid syndrome | 1 | 1000 | 0 | 1000 |
| Total | | 394,827 | 35,091 | 359,838 |

[a] No data (series > 1000) was found for the following SADs: relapsing polychondritis, Still disease, hemophagocytic syndrome, IgG4-related disease, polyarteritis nodosa, Takayasu arteritis, Buerger disease, eosinophilic granulomatosis with polyangiitis, microscopic polyangiitis and Cogan disease.