LUPUS SCIENCE & MEDICINE

Dialogue: High-throughput studies in rheumatology: time for unsupervised clustering?

George Bertsias [1,2]

To cite: Bertsias G. Dialogue: High-throughput studies in rheumatology: time for unsupervised clustering?. *Lupus Science & Medicine* 2021;**8**:e000643. doi:10.1136/lupus-2021-000643

Received 15 December 2021
Accepted 15 December 2021

[1]Rheumatology, Clinical Immunology and Allergy, University of Crete School of Medicine, Heraklion, Crete, Greece
[2]Laboratory of Rheumatology, Autoimmunity and Inflammation, Institute of Molecular Biology and Biotechnology (IMBB-FORTH), Heraklion, Greece

Correspondence to
Dr George Bertsias; gbertsias@uoc.gr

In complex autoimmune rheumatic diseases, high-throughput technologies simultaneously analysing dozens, hundreds or thousands of biological cues (genes, metabolites, serum proteins etc) have long been considered valuable in obtaining unique pathogenic insights while facilitating the discovery of therapeutic targets and biomarkers for diagnosis, monitoring and prognosis.[1]

In the current issue of *Lupus Science and Medicine*, Brunekreef *et al*[2] used a custom chip-based microarray to probe serum samples for a total 57 known and new IgG autoantibodies and explore their diagnostic utility in SLE. By comparing the prevalence of each autoantibody in 483 patients with SLE and 1397 disease controls (including 361 healthy individuals), they found that anti-double stranded(ds)DNA antibodies and antibodies against Cytosine-phosphate-Guanine (anti-CpG) DNA motifs could best discriminate SLE versus control groups with corresponding area under the receiver operating curve (AUC) values of 0.800 and 0.756, respectively.[2] Notably, 15.1% of patients with SLE negative for anti-dsDNA tested positive for anti-CpG DNA antibodies, therefore suggesting added diagnostic value. Although the exact specificity of CpG-targeting antibodies was not explored and some cross-reactivity with anti-dsDNA antibodies cannot be entirely excluded, the results are biologically plausible given the abundance of nucleic acids containing unmethylated or hypomethylated CpG DNA in SLE.[3–5]

Pending further standardisation of the CpG DNA detection methods and validation of these findings, certain methodological aspects of this work merit discussion. First, patients were designated as SLE or other disease/condition by the use of a *text mining algorithm* that searched for pre-specified disease-related or symptom-related keywords in retrospectively collected electronic health records. Although, in general, such strategies are considered valid and advantageous for large datasets,[6] algorithm-assigned diagnoses were not ascertained by the existing classification criteria or other means. This might account for the lower-than-expected frequency of antinuclear antibodies (19 out of 147 first samples tested negative) in patients with SLE and also the fact that about 30% of all patients received more than one diagnosis.

Second, the researchers assigned patients without SLE to multiple control groups including one with mild, non-specific symptoms resembling healthy controls, a second with lupus-like (or incomplete lupus) presentations (eg, arthritis, nephritis, serositis) and a third with an autoimmune disease other than SLE.[2] Notwithstanding this might reflect the 'real-life' situation where patients do not always fit into exact diagnostic entities, one should consider that autoimmune rheumatic diseases like SLE tend often to develop over time; therefore, some of the disease controls might represent early (or pre-) lupus forms.[7 8] This is also supported by the between-group differences in the prevalence of autoantibodies reported by the authors.[2]

These complexities in the definition and phenotypic heterogeneity of autoimmune rheumatic disorders bring out the issue of how we can best use high-throughput studies and big data towards disease diagnosis/classification and risk stratification. To date, the majority of studies have employed a conventional, 'supervised'-type approach to analyse biological (*input*) data which are tagged with pre-specified (*output*) 'labels' (diagnostic or endophenotypic groups). Although this method is straightforward and can yield accurate classification results, especially following implementation of sophisticated machine learning tools,[9–11] it is biased heavily on the accuracy of the available diagnostic information (considered to be '*ground truth*') and pre-existing grouping of the dataset. In the situation we have no accurate prior knowledge on the diagnostic groups for the samples or the output is not really "*yes or no*" (eg, SLE or not) but rather behaves as a continuum of states (eg, ranging from healthy, pre-lupus, mild lupus, severe lupus), *unsupervised clustering* (or *learning*) might represent a more suitable solution.

Indeed, these computational methods require no preconceived assumptions, work with unlabeled outputs and infer the inherent structure present within a dataset.[10 12] Accordingly, they are useful to recognise hidden patterns or combinations of biological data, therefore providing a natural clustering of the complex-structured samples. Interpretability of the resulting clusters and characterisation of their distinctive features in a compact form may require additional steps as part of a decision-making process;[13] nonetheless, unsupervised approaches move closer to the current concept of revisiting autoimmune rheumatic diseases based on the underlying molecular taxonomy.[14]

To this end, high-throughput studies such as this by Brunekreef *et al*[2] represent notable contributions in the diagnostics of rheumatic diseases and the identification of sub-phenotypes with possibly distinct underlying pathophysiology. With accruing experience in the analysis of big data, the community should gradually move forward to implementing less biased classification methods to ultimately '*let the data speak for themselves*'.

Contributors I am the sole contributor of this article (Editorial).

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD

George Bertsias http://orcid.org/0000-0001-5299-1406

REFERENCES

1 Donlin LT, Park S-H, Giannopoulou E, *et al*. Insights into rheumatic diseases from next-generation sequencing. *Nat Rev Rheumatol* 2019;15:327–39.
2 Brunekreef T, Limper M, Melchers R, *et al*. Microarray testing in patients with systemic lupus erythematosus identifies a high prevalence of CpG DNA-binding antibodies. *Lupus Sci Med* 2021;8:e000531.
3 Anders HJ. A toll for lupus. *Lupus* 2005;14:417–22.
4 Caielli S, Athale S, Domic B, *et al*. Oxidized mitochondrial nucleoids released by neutrophils drive type I interferon production in human lupus. *J Exp Med* 2016;213:697–713.
5 Vecellio M, Wu H, Lu Q, *et al*. The multifaceted functional role of DNA methylation in immune-mediated rheumatic diseases. *Clin Rheumatol* 2021;40:459–76.
6 Sun W, Cai Z, Li Y, *et al*. Data processing and text mining technologies on electronic medical records: a review. *J Healthc Eng* 2018;2018:1–9.
7 Bourn R, James JA. Preclinical lupus. *Curr Opin Rheumatol* 2015;27:433–9.
8 Mankia K, Emery P. Preclinical rheumatoid arthritis: progress toward prevention. *Arthritis Rheumatol* 2016;68:779–88.
9 Adamichou C, Genitsaridi I, Nikolopoulos D, *et al*. Lupus or not? SLE risk probability index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Ann Rheum Dis* 2021;80:758–66.
10 Jiang M, Li Y, Jiang C, *et al*. Machine learning in rheumatic diseases. *Clin Rev Allergy Immunol* 2021;60:96–110.
11 Kingsmore KM, Puglisi CE, Grammer AC, *et al*. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nat Rev Rheumatol* 2021;17:710–30.
12 Sohail A, Arif F. Supervised and unsupervised algorithms for bioinformatics and data science. *Prog Biophys Mol Biol* 2020;151:14–22.
13 Bertsimas D, Orfanoudaki A, Wiberg H, 2021. Available: https://dbertsim.mit.edu/pdfs/papers/2021-wiberg-interpretable-clustering-an-optimization-approach.pdf
14 Barturen G, Beretta L, Cervera R, *et al*. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nat Rev Rheumatol* 2018;14:75–93.