Response to: 'On using machine learning algorithms to define clinically meaningful patient subgroups' by Pinal-Fernandez and Mammen

We have read with interest the comment from Pinal-Fernandez and Mammen in which they question the statistical clustering methods based on unsupervised learning analyses to define clinically meaningful patient subgroups.[1] Pinal-Fernandez and Mammen base their arguments on the production of an analysis according to this methodology made on a random simulated data set that would highlight the formation of three clusters, in fact arbitrary.

It is important to point out that the example which forms the basis of their argument is ill-chosen because it shows a misguided use of this type of technique. Indeed, before applying a clustering method to a data set, good practice recommendations must be followed.[2]

- First of all as with any experiment, one should ask the question of the clinical and/or scientific relevance of the research. Obviously, wanting to classify a completely random simulated data set has no interest. On the other hand, if we take the case of myositides, proposing an intra-syndromic classification justified by 50 years of medical literature and debate[3] is obviously much more relevant.
- Since clinical relevance may not always be obvious, there are statistical tools that make it possible to judge whether statistical groupings are appropriate.[2] Bartlett's spherical test proposes an overall measure based on a statistical approach. This test will not predict the existence of an interesting partition but at least, it will indicate whether it seems appropriate to aggregate this information. Here if we use the simulation experiment proposed by Pinal-Fernandez and Mammen on a large number of times (eg, 10 000), Bartlett's test will not reject the null hypothesis in 95% of cases, and therefore the reason would have led them not to apply a clustering method on this data set.

Nevertheless, the comment of Pinal-Fernandez and Mammen has the merit to highlight the very real problem of the optimal number of clusters, underpinned by the fact that there is no straightforward definition of what a cluster is. This problem is well known as the elbow phenomenon and attempts to deal with it are well documented.[4] The general principle of selecting this optimal cluster number is to measure a classification error and calculate it in relation to the proposed number of clusters. There are 'global' methods which determine the total performance of the classification and the so-called 'local' methods which work on cluster pairs and which make it possible to judge whether they are justified. The idea here is not to compile a catalogue of these different methods and their advantages and disadvantages but rather to say that a user accustomed to these methods knows this problem and has a large number of tools to apprehend it.[5]

Olivier Benveniste ,[1] **Yves Allenbach,**[1] **Benjamin Granger**[2]

[1]Department of Internal Medicine and Clinical Immunology and Paris Neuromuscular Rare Diseases Reference Center, Sorbonne Université, INSERM U974, Assistance Publique-Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Paris, France
[2]Department of Biostatistics and Clinical Information, Sorbonne Université, INSERM UMR 1136, Assistance Publique-Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Paris, France

Correspondence to Dr Olivier Benveniste, Department of Internal Medicine and Clinical Immunology, Hospital University Department: Inflammation, Immunopathology and Biotherapy (DHU i2B), Assistance Publique - Hôpitaux de Paris, Pitié-Salpêtrière University Hospital, Paris 75013, France; olivier.benveniste@aphp.fr

Handling editor Josef S Smolen

Contributors All authors wrote this reply together.

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 consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

To cite Benveniste O, Allenbach Y, Granger B. *Ann Rheum Dis* 2020;**79**:e130.

Received 10 July 2019
Revised 10 July 2019
Accepted 12 July 2019
Published Online First 20 July 2019

► http://dx.doi.org/10.1136/annrheumdis-2019-215852

Ann Rheum Dis 2020;**79**:e130. doi:10.1136/annrheumdis-2019-216007

ORCID iD
Olivier Benveniste http://orcid.org/0000-0002-1167-5797

REFERENCES

1 Pinal-Fernandez I, Mammen AL. On using machine learning algorithms to define clinically meaningful patient subgroups. *Ann Rheum Dis* 2020;79:e128.
2 Williams B, Onsman A, Brown T. Exploratory factor analysis: a five-step guide for novices. *Australas J Paramed* 2010;8.
3 Mariampillai K, Granger B, Amelin D, *et al*. Development of a new classification system for idiopathic inflammatory myopathies based on clinical manifestations and myositis-specific autoantibodies. *JAMA Neurol* 2018;75:1528–37.
4 Gordon AD. *Classification*. 2nd edition. London: Chapman and Hall, 1999.
5 Yan M, Ye K. Determining the number of clusters using the weighted gap statistic. *Biometrics* 2007;63:1031–7.