On how to not misuse hierarchical clustering on principal components to define clinically meaningful patient subgroups. Response to: 'On using machine learning algorithms to define clinical meaningful patient subgroups' by Pinal-Fernandez and Mammen

We thank Pinal-Fernandez and Mammen for their interesting methodological comment on our work in which we used hierarchical clustering on principal components to define clinically meaningful subgroups of patients with anti-Ku antibodies.^{1 2}

We fully agree with the conclusion of the authors: 'machine learning methods may be fundamentally flawed if a cornerstone of the analysis depends upon the incorrect use of a complex biostatistical technique'.

In this regard, the example of hierarchical clustering on principal components they provide in their comment is an illustration on how this statistical tool can be misused and generate false discoveries:

1. First, hierarchical clustering on principal components is a descriptive method that is fitted to describe heterogeneous datasets. Prior reports from the literature indicate that patients with anti-Ku did represent a heterogeneous condition. On the contrary, in their example, Pinal-Fernandez and Mammen (voluntarily) used a multivariate normal distribution that is perfectly unfitted for hierarchical clustering on principal components.
2. Second, the suitability of the dataset for clustering analysis on principal components can be further tested statistically, for example, using the Bartlett's test of sphericity as performed in our study. Bartlett's test is an objective benchmark which is used to minimise the danger of interpreting factor analytic results which can be attributed entirely to chance.³ In accordance with the first point, in our dataset from anti-Ku patients, the p value was <0.001, indicating that hierarchical clustering on principal components was very likely to be useful. On the contrary, we ran the Bartlett's test of sphericity on 1000 datasets proposed by Pinal-Fernandez and Mammen and found that the p value was >0.05 in 95% of cases (which was expected since variables were simulated independently).
3. Third, principal components analysis (PCA) enables to distinguish structure from noise contained in the data. The number of dimensions retained to determine the principal components of the dataset is a crucial step of the analysis.⁴ In this perspective, we used K-fold cross-validation and retained the number of dimensions leading to the smallest mean square error of prediction (MSEP). On the contrary, in their example, Pinal-Fernandez and Mammen arbitrarily fixed the number of dimensions to two. Yet, in their dataset, PCA produces four components which equally explain ~25% of the variance of the data because there is no structure in the dataset. Thus, the K-fold cross-validation applied on their dataset indicated that the number of dimensions leading to the smallest MSEP is zero, which means that there

is no principal component in their dataset, but only noise.

In such case, hierarchical clustering must not be undertaken.

As illustrated by the more than 5000 original articles referenced on PubMed using this method, hierarchical clustering does represent a powerful tool to describe datasets.

As smartly illustrated by Pinal-Fernandez and Mammen, several rules must be followed in order to not misuse hierarchical clustering on principal components.

Alain Meyer,^{1,2} **Lionel Spielmann** ,³ **François Séverac**^{4,5}

¹Exploration Fonctionnelle Musculaire, Service de physiologie, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

²Centre National de Référence des Maladies Auto-Immunes Systémiques Rares de l'Est et du Sud-Ouest, Service de rhumatologie, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

³Service de Rhumatologie, Hôpitaux Civils de Colmar, Colmar, France

⁴Service de Santé Publique, GMRC, CHU de Strasbourg, Strasbourg, France

⁵iCUBE, UMR 7357, équipe IMAGEs, Université de Strasbourg, Strasbourg, France

Correspondence to Dr Lionel Spielmann, Service de Rhumatologie, Hospices Civils de Colmar, Colmar 68024, Alsace (Région), France; lionel.spielmann@ch-colmar.fr

Correction notice This article has been corrected since it published Online First. The title has been amended.

Handling editor Prof Josef S Smolen

Contributors AM supervised the project. AM and LS wrote the manuscript with support from FS. FS verified the analytical method. AM, LS and FS contributed to the final version of the manuscript.

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 No, there are no competing interests for any author.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

© Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ.

To cite Meyer A, Spielmann L, Séverac F. *Ann Rheum Dis* 2020;**79**:e129.

Received 8 July 2019

Accepted 9 July 2019

Published Online First 24 July 2019

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

Ann Rheum Dis 2020;**79**:e129. doi:10.1136/annrheumdis-2019-215868

ORCID iD

Lionel Spielmann <http://orcid.org/0000-0003-1057-6890>

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 Spielmann L, Nespola B, Séverac F, et al. Anti-Ku syndrome with elevated CK and anti-Ku syndrome with anti-dsDNA are two distinct entities with different outcomes. *Ann Rheum Dis* 2019;**78**:1101–6.
- 3 Tobias S, Carlson JE. Brief report: BARTLETT's test of SPHERICITY and chance findings in factor analysis. *Multivariate Behav Res* 1969;**4**:375–7.
- 4 Josse J, Husson F. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput Stat Data Anal* 2012;**56**:1869–79.