On using machine learning algorithms to define clinically meaningful patient subgroups

Improved taxonomy will drive our efforts to personalise medicine over time. Ideally improved taxonomy is fueled by our detailed insight in pathogenesis leading to subgrouping syndrome's into more homogeneous diseases. An alternative is to cluster subgroups of patients based on similar manifestations and prognosis. So, the recent publication of Spielman *et al*[1] as well as the correspondence on that study written by Pinal-Fernandez and Mammen[2] is very timely and interesting. Spielman *et al*[1] identified three clinical clusters in patient with anti-Ku-positive myositis by applying hierarchical clustering analysis on both clinical and biological features.

Pinal-Fernandez and Mammen suggest that the results of Spielman's work might be flawed as they disagree with the method of number of cluster selection. Indirectly they also challenge the idea of using (hierarchical) clustering techniques to identify clinically meaningful patient populations. Of course, we agree that improper use of analytical methods can lead to incorrect conclusions. Therefore, we applaud the ongoing discussion on how to reliably use 'big-data techniques' partly fueled by EULAR's point to consider for the use of big data (techniques).[3] To contribute to this discussion, we would like to challenge the statements made by Pinal-Fernandez and Mammen.[2]

IDENTIFICATION OF THE NUMBER OF CLUSTERS

In order to find the 'optimal' number of clusters, several methods can be employed. The elbow method (mentioned in the correspondence) ranks clusters by the variance they explain and defines the cut-off at the point where additional clusters do not substantially increase the explained variance.

A different method that has a higher face validity is the consensus clustering.[4] Here, the clustering analysis is run on different subsets of the data and the proportion in which samples cluster together in all attempts is depicted in a heatmap. In case of a good consensus, the heatmap depicts the anticipated number of correlation blocks.

Alternatively, one can incorporate clinical knowledge to define the anticipated number of clusters. The optimal number depends on the available meaningful consequences of the cluster identification: for example, number of treatment options, number of different long-term outcomes, anticipated aetiologic differences and so on. Discovering a few more clusters than, for instance, treatment options is interesting in the light of new drug discovery, but identification of even more clusters will probably have little scientific value.

Pinal-Fernandez and Mammen already stated that there is no optimal way of finding the 'right' number of clusters. Their simulation does not demonstrate that the findings of Spielman *et al*[1] are flawed. In contrast, they demonstrate the importance of validation of the results, because indeed, clustering methods have the ability to identify patterns even in data with very little structure.

VALIDATION OF RESULTS

Consequently, the most relevant question is how to validate results from 'big-data' or machine learning methods. First of all, there are the traditional methods of validation: replication in a second cohort, replication in an untouched second part of the original dataset and correlation of the identified groups with (long-term serious) complications. When additional cohorts are not at hand and the available dataset is too small to divide into a training and a test dataset, one could use *k*-fold cross validation (CV). CV reruns an analysis multiple (*k*) times on a slightly different version of the data. The resulting information can be used to adjust the classification model such that it does not overfit or it can be used to describe the precision of the results.

Spielman *et al*[1] used 1000-fold CV to prevent overfitting of their dimensionality reduction. Second, they further studied their identified clusters and showed that there is a clear difference in serious disease complications (interstitial lung disease (ILD) and glomerulonephritis). Subsequently, they identified baseline elevated creatine kinase and anti-dsDNA as important variables for the prediction of ILD and glomerulonephritis with substantial effect estimates (13-fold and 22-fold risk for some patients). This cluster validation step would have been more convincing if ILD and glomerulonephritis were not included in the cluster identification, but still the results are convincing.

Pinal-Fernandez and Mammen[2] are correct in noting that reaching significance in data of >1000 samples is inevitable, but Spielman *et al*[1] only had 42 patients in their study and their p values survive Bonferroni correction for multiple testing (which corrects for the number of performed tests n=40).

VALUE OF USING CLUSTERING TECHNIQUES FOR PATIENT GROUP IDENTIFICATION

The use of clustering methods (most of which can be categorised under the machine learning methods) could proof useful for our field of research where most of the studied diseases are complex and the clinical presentation and outcomes are heterogeneous.[5] When diseases are too complex and too rare for us to identify homogeneous patient groups in clinic or with one-on-one associations, clustering techniques are interesting analytical tools.

Therefore, we conclude that machine learning methods offer great opportunities and we embrace their availability for both statistical experts and non-experts. These methods require a different approach to research then we were used to: it is not about choosing the one and only correct method since many machine learning methods are equally useful. Instead, it is crucial to validate the findings. So, in the evaluation of results, we suggest focusing on the validation of the research results.

Rachel Knevel , **Tom WJ Huizinga**

Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence to Dr Rachel Knevel, Department of Rheumatology, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands; r.knevel@lumc.nl

Contributors The authors wrote the paper together.

Funding This study was supported by Reumafonds.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

To cite Knevel R, Huizinga TWJ. *Ann Rheum Dis* 2020;**79**:e154.

Received 28 June 2019
Accepted 4 July 2019
Published Online First 11 July 2019

Linked

► http://dx.doi.org/10.1136/annrheumdis-2019-215989

Ann Rheum Dis 2020;**79**:e154. doi:10.1136/annrheumdis-2019-215959

ORCID iD
Rachel Knevel http://orcid.org/0000-0002-7494-3023

REFERENCES

1 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.
2 Pinal-Fernandez I, Mammen AL. On using machine learning algorithms to define clinically meaningful patient subgroups. *Ann Rheum Dis* 2020;79:e128.
3 Gossec L, Kedra J, Servy H, *et al*. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. *Ann Rheum Dis* 2020;79:69–76.
4 Monti S*et al*. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 2003;52:91–118.
5 Seymour CW, Kennedy JN, Wang S, *et al*. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019;321.