Prof. Dr. Matthias Kohl
Institute of Precision Medicine
Furtwangen University

The

Response Letter to Vink et al. 'Neurological Study Does Not Provide Any Evidence that Long COVID is Psychosomatic' by Stettner et al. (2022)

is very unsatisfactory with respect to the problems raised by us in general and with respect to the statistical weaknesses in particular.

"Finally, several issues related to our statistical analysis were raised. The authors assume 'a high probability of multiple false positives'. The rationale behind this conclusion remains unclear."

The article includes the results of many statistical tests. For instance in Fig. 2 the results of 3 x 8 = 24 statistical tests are shown (panel A, B, C). Since the chance that some of the results are false positive increases with the number of tests, we would like to adjust the p values by the method of Holm hence controlling the family wise error rate (FWER). For doing this, we would actually need the exact p values, which are however not given in the article (see also below). Therefore, we will argue with the given p value categories. Four of the significant results are in the range of $0.01 < p < 0.05$. In case of the method of Holm the smallest p value must be multiplied with the number of tests (24). Even if we would consider the tests panel wise (multiplication by 8), the corresponding adjusted p values would be larger than 0.05. Hence, only two test results remain that might be significant after the adjustment, which are the tests with results in the range $0.001 < p < 0.01$. Since the exact p values are not shown, it is unfortunately unclear whether a significant result would remain after adjusting these two p values either panel wise or for the whole Fig. 2.

Similar considerations also apply to the other test results given in the article. Overall this shows that there is a clear multiple testing issue and it is unclear, how many tests would still be significant after adjusting for multiple testing and hence controlling the FWER.

In addition, we could not find any information about the registration of the study, which is a standard requirement also for observational studies (see https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html). Therefore, it is unclear which and how many statistical analyses were originally planned. Moreover, it is also unclear, whether all planned analyses were also reported and whether new analyses were added after the start of the study.

Therefore, one should consider the study only as explorative and the results need to be confirmed by new independent studies.

"They claim that for the stepwise regression analyses '*p*-values and standard errors are too small and confidence intervals too short'. This statement is inherently linked to the stepwise regression approach and is not a specific limitation of our study."

It is true that it is a weakness of the stepwise approach. However, we don't understand why this approach was used by the authors anyway. There are also several other approaches for variable/feature selection that could have been used instead.

"As a matter of fact, the results stand as they are and were confirmed by the statistical tests described in the Methods section."

In view of the weaknesses described above, we think “confirmed” is a much too strong word here. We would consider the study only as exploratory, hence the confirmation of the results is still pending.

“Moreover, our manuscript underwent extensive peer review, including statistics and editorial review, prior to publication.”

We have some doubts regarding a statistical review, since the article includes several weaknesses and imprecise statements regarding the statistical analysis.

“Also, the reporting of statistical findings was in full accordance with the Journal’s submission guidelines, which do not require reporting of *p*-values to two or three decimal places [11].”

As demonstrated above, we would strongly recommend to update the statistical guidelines. They are no longer representing the state of the art (see for instance https://www.nejm.org/author-center/new-manuscripts).

“Regarding the clustering approach, categorical variables of complaints were used in a two-step algorithm, which is accepted as a robust test. For the multinomial distribution of the variants, the log-likelihood measure of distance was performed. We can confirm the verification of underlying distributional assumptions with the chosen variables' independence by using the χ^2-test. In addition, order dependence was fully considered during the statistical analysis by clustering the variables multiple times in different orders.”

Since the authors state that they have repeated the clustering multiple times, we ask ourselves which of these results was selected for the final article and based on what criteria. In addition, the article includes no word about the stability of the clusters, which would be an important result of such an analysis.