Concerns Regarding Linear Assumptions in Principal Component Analysis: Advocating for Nonlinear and Nonparametric Approaches in Pulmonary Carcinoid Research

To the Editor:

Leunissen et al.[1] identified defined molecular subgroups on the basis of immunohistochemical analyses and potential therapeutic vulnerabilities of pulmonary carcinoids. They reported that unsupervised clustering was performed using the k-means method, followed by dimensionality reduction via principal component analysis (PCA). Their analysis utilized an enriched dataset of protein expression, including OTP, HNF1A, ASCL1, S100, and TTF1, correlating these proteins with various clinical characteristics. In essence, PCA was employed for feature reduction and to assess feature importance.[1]

Nevertheless, this article raises significant concerns regarding the application of PCA for feature reduction and importance assessment owing to its linear and parametric nature. Although Leunissen et al.[1] are knowledgeable in thoracic oncology, they may lack expertise in algorithmic calculations, which can introduce biases and lead to erroneous conclusions. PCA's inherent linear assumptions necessitate careful consideration of whether to employ linear versus nonlinear and parametric versus nonparametric approaches. Without this critical evaluation, essential features may be overlooked, resulting in substantial biases. Therefore, it is crucial for researchers to consider nonlinear and nonparametric methods to improve the robustness of their analyses and reduce potential biases.

This article advocates for adopting bias-free robust statistical methods such as Spearman's correlation with p values or Kendall's tau with p values, both nonlinear and nonparametric approaches.[2–4]

Imposing linear and parametric assumptions on nonlinear data can lead to severe biases for several reasons.[5–8] First, linear models fit an assumption of a straight-line relationship to the data. When the actual relationship is nonlinear, the predictions made by the model can significantly deviate from reality, resulting in biased estimates and conclusions. In addition, this inability to capture the complexities of nonlinear patterns often leads to underfitting, which can compromise the predictive performance and misguide interpretations of the data. Moreover, parametric approaches commonly assume that data follow a specific distribution—such as a normal distribution—and when these assumptions are violated, standard statistical inference techniques may yield incorrect results, raising the potential for erroneous conclusions. Furthermore, important features or relationships may be overlooked when applying linear methods, as these techniques fail to appropriately represent the underlying structure of the data. Consequently, significant variables could be missed, which may lead to misguided strategies or policies on the basis of incomplete or inaccurate information. Collectively, these factors underscore the importance of employing appropriate modeling techniques that align with the underlying distribution and structure of the data, ensuring that analyses yield valid and reliable insights.

CRediT Authorship Contribution Statement

Yoshiyasu Takefuji: Writing - original draft, Writing - review & editing.

Yoshiyasu Takefuji, PhD
Faculty of Data Science
Musashino University
Tokyo Japan

Disclosure

The authors declare no conflict of interest.

References

1. Leunissen DJG, Moonen L, von der Thüsen JH, et al. Identification of defined molecular subgroups on the basis of immunohistochemical analyses and potential therapeutic vulnerabilities of pulmonary carcinoids. *J Thorac Oncol*. 2025;20:451–464.
2. Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. *Commun Stat Theory Methods*. 2024;53:2141–2153.
3. Eden SK, Li C, Shepherd BE. Nonparametric estimation of Spearman's rank correlation with bivariate survival data. *Biometrics*. 2022;78:421–434.
4. Wang JH, Chen YH. Network-adjusted Kendall's tau measure for feature screening with application to high-dimensional survival genomic data. *Bioinformatics*. 2021;37:2150–2156.

Address correspondence to: Yoshiyasu Takefuji, PhD, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan. E-mail: takefuji@keio.jp

Cite this article as: Takefuji Y. Concerns regarding linear assumptions in principal component analysis: advocating for nonlinear and nonparametric approaches in pulmonary carcinoid research. *J Thorac Oncol* 2025;20:e54-e55.

ISSN: 1556-0864
https://doi.org/10.1016/j.jtho.2024.12.024

5. Rocks JW, Mehta P. Memorizing without overfitting: bias, variance, and interpolation in overparameterized models. *Phys Rev Res*. 2022;4:013201.
6. Montoye AHK, Begum M, Henning Z, Pfeiffer KA. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiol Meas*. 2017;38:343–357.
7. Domingue BW, Kanopka K, Trejo S, Rhemtulla M, Tucker-Drob EM. Ubiquitous bias and false discovery due to model misspecification in analysis of statistical interactions: the role of the outcome's distribution and metric properties. *Psychol Methods*. 2024;29:1164–1179.
8. Lazic S. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed.

A Response to the Letter to the Editor: "Linear Versus Non-Linear: Debunking Critiques on PCA Use in Molecular Subgrouping of Pulmonary Carcinoids"

To the Editor:

We have read the letter from Dr. Takefuji[1] entitled "Concerns Regarding Linear Assumptions in PCA: Advocating for Nonlinear and Nonparametric Approaches in Pulmonary Carcinoid Research," which we received in response to our article on the identification of defined molecular subgroups on the basis of immunohistochemical analyses of pulmonary carcinoids.[2]

In our article, we reported that unsupervised clustering was performed using the k-means algorithm, followed by dimensionality reduction using principal component analysis (PCA) employed as a verification method. Our analysis utilized an enriched data set of protein expression, including OTP, HNF1A, ASCL1, S100, and TTF1, in correlation with various clinical characteristics.[2] Takefuji[1] raises concerns regarding the application of PCA for feature reduction and the assessment of feature importance due to its linear nature, which might result in potential biases in data interpretation. Nevertheless, it needs to be emphasized that PCA was not employed for feature reduction nor for the assessment of feature importance within our research as the author mistakenly points out, and that the author overlooked that we actually did perform the very nonlinear nonparametric analyses (Spearman correlations) that he proposes.

Address correspondence to: J.L. Derks, MD, PhD, Department of Pulmonary Medicine, Erasmus MC Cancer Institute, University Medical Center, Dr. Molewaterplein 40, 3015 GD Rotterdam, Postbox 2040, Rotterdam 3000 CA, The Netherlands. E-mail: j.derks@erasmusmc.nl

Cite this article as: Leunissen DJG, Moonen L, Alcala N, et al. A response to the letter to the editor: "Linear Versus Non-Linear: Debunking Critiques on PCA use in molecular subgrouping of pulmonary carcinoids." *J Thorac Oncol*. 2025;20:e55-e56.

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2025.01.012

Our clinical cohort was divided into molecular subgroups on the basis of the immunohistochemical marker H-scores, which is a form of rank-based method, not assuming linearity. The H-score thresholds were defined using an independent matched RNA and protein cohort. Also note that these RNA-defined molecular subgroups have been extensively benchmarked, independently of our study, in the study of Gabriel et al.,[3] in which the performance of PCA (in particular, its ability to preserve the nearest-neighbors from the original high-dimensional space) was extensively compared with that of UMAP, which is a nonlinear graph-based method. UMAP was also performed by Dayton et al.,[4] who yielded similar groupings. After subgrouping of our clinical cohort, PCA was solely employed to verify an agreement between the immunohistochemical-RNA defined clusters and the PCA-formed clusters. As reported in the article, this analysis resulted in a 94% agreement between the immunohistochemical-RNA defined clusters and the PCA-formed clusters, proving the agreement between linear and non-linear clustering methods. Furthermore, PCA clustering was performed on a set of unclassifiable cases to understand the differences of these cases compared with the rest of the cohort. This PCA was performed to investigate which clusters would form and to determine the molecular subgroup to which they most closely correspond. In addition, Spearman correlations were indeed performed among all different marker combinations as defined within the supplementary data (Supplementary Fig. 3). Finally, note that the data presented have been established in close collaboration with biostatisticians and bioinformaticians (i.e., Leunissen, Alcala, Foll) with knowledgeable insight into applications of statistical methodology. Therefore, the suggestion of a lack of expertise in algorithmic calculation seems too strongly worded.

In conclusion, the suggestions made by Dr. Takefuji seem incorrect and a more careful appreciation of the (supplementary) data would have provided these insights. Nevertheless, we support his statement on the importance of the correct application of algorithmic calculations in (thoracic) research.

CRediT Authorship Contribution Statement

Daphne J.G. Leunissen: Conceptualization, Writing - original draft.

Laura Moonen: Writing - review & editing.