Non-invasive Diagnosis of Prostate Cancer via Urine: A multivariate model based on LC-HRMS and PLS-DA

Lucas G. Costa (PG)[1]*, Waldomiro Borges Neto (PQ)[1]

[1] Universidade Federal de Uberlândia, Instituto de Química, Uberlândia/MG
*lucasgustavodacosta@ufu.br

RESUMO

O câncer de próstata é a segunda neoplasia mais comum entre homens no Brasil e no mundo. Sua detecção precoce é essencial para o sucesso do tratamento, mas os métodos atuais apresentam limitações. Assim, este estudo propõe um método não invasivo para sua detecção através da análise de urina, combinando cromatografia líquida acoplada a espectrometria de massas de alta resolução com a ferramenta multivariada Análise Discriminante por Mínimos Quadrados Parciais. Foram analisadas amostras de 40 pacientes com câncer e 40 saudáveis, obtendo sensibilidade e especificidade de 100 %, sem falsos positivos ou negativos. O modelo demonstrou robustez e permitiu a identificação de picos espectrais relevantes como potenciais biomarcadores. Essa abordagem apresenta-se como uma alternativa promissora para diagnósticos precisos e menos invasivos quando comparada aos métodos tradicionais, porém, requer validação em conjuntos de amostras maiores.

Palavras-chave: Non-invasive diagnostics, Biomarkers, Metabolomics, Chemometrics.

Introduction

Prostate cancer (PC) is the second most common neoplasm among men in Brazil, surpassed only by non-melanoma skin cancer when considering both sexes. Globally, PC is the fourth most common cancer overall and the second most common among men (1), indicating the seriousness of the disease. Early detection is essential for successful treatment; however, the prostate-specific antigen (PSA) and digital rectal examination methods are limited in terms of sensitivity and specificity, and they are uncomfortable (2).

In this context, urine-based diagnoses have great potential because urine is easy to collect, noninvasive, and rich in metabolic information that can reflect the patient's physiological state. However, due to the complexity of this biological matrix, multivariate analysis tools are necessary to extract as much information as possible and create robust discrimination models. This study developed and validated an analytical model based on liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS), combined with the chemometric method of partial least squares discriminant analysis (PLS-DA), to separate healthy individuals from those with prostate cancer through urine.

Experimental

The model was created using urine samples from 40 patients diagnosed with PC and 40 healthy individuals with no previous history of the disease. Samples from the PC group were obtained from the Prostate Cancer Biospecimen Core of the SPORE network at Northwestern University in the USA, and samples from the healthy group were obtained from the Life Study at the University of Florida in the USA. Approval was obtained from the institutions' Ethics Committees. After collection, the samples were stored in cryogenic tubes (Norgen Biotek, Canada) at -80 °C.

Spectra were obtained by LC-HRMS using a Dionex Ultimate 3000 system and a Q Exactive spectrometer (Thermo Scientific). Ionization was conducted in negative mode, covering the range from 70 to 1000 m/z with a resolution of 35,000. For chromatographic separation, an ACE Excel 2 C18-PFP column (100 × 2.1 mm, 2.0 µm) was used with a mobile phase of water containing 0.1 % formic acid (solvent A) and acetonitrile (solvent B) at a flow rate of 350 µL/min. The elution gradient was carried out from 100 % solvent A to 20 % in 10 min, followed by 3 min at 20 %, then returning to the initial conditions in 0.5 min. The column temperature was maintained at 25 °C.

The spectral data were organized into a matrix and subjected to preprocessing, which consisted of normalizing the peaks according to the highest detected signal and removing peaks with a relative abundance of less than 5 %. The dataset was then divided into 52 samples for calibration (26 PC and 26 healthy) and 28 samples for testing (14 PC and 14 healthy). We carried out chemometric modeling using MATLAB software (version R2021a) with PLS_Toolbox (version 9.5). We applied the mean center of the variables and the Venetian blinds cross-validation method (12 windows, each with 5 samples) to classify the samples. We assessed the model's performance using the following parameters: sensitivity, specificity, Matthew's correlation coefficient, and root mean square error of calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP).

Results and Discussion

The PLS-DA model built with five latent variables (VL) showed satisfactory statistical performance, as evidenced by the low error values: RMSEC of 0.53 %, RMSECV of 0.64 % and RMSEP of 0.58 %. These results suggest that the model fits the data well in both the calibration and cross-validation stages, demonstrating the

model's robustness and stability. The slightly higher RMSEP than RMSEC reinforces the correct fit and shows that there were no underfitting or overfitting issues.

The model's sensitivity and specificity were both 100%, showing the method's ability to correctly identify all samples in the two evaluated classes, with no false positives or negatives in the test set. Consequently, the Matthews correlation coefficient was +1, reinforcing the distinction between the PC group and the healthy group. These results are promising, especially considering the difficulty of discrimination in complex biological matrices, such as urine, where metabolic profiles vary greatly between individuals.

With a class separation threshold of -0.04, the PLS-DA model's estimation graph clearly illustrates the separation between the groups. There is a clear distribution of samples above and below the threshold that defines classification as either patients with CP or healthy individuals. This can be easily interpreted by anyone without the need for specific knowledge, as can be seen in Figure 1.

In addition to its high classification capacity, the model identified the most relevant variables (peaks in the mass spectra) for separating the classes by analyzing the variables important to the projection (VIP). The most relevant peaks identified in this study can serve as a starting point for future structural elucidation studies, clinical validation, and potentially, the development of new diagnostic tests. Identification of these peaks makes it possible to find potential prostate cancer biomarkers, which could increase the efficiency and simplify future tests.

Conclusion

The combined use of high-resolution mass spectrometry and partial least squares discriminant analysis shows promise for screening prostate cancer in urine samples. The model's high predictive capacity offers an initial alternative for developing more accurate, less invasive, and efficient diagnostic methods.
While the number of samples is sufficient to create the initial model, it is limited for broad clinical applicability. Further studies with larger sample sizes, covering different populations and clinical conditions, are necessary to validate the method and confirm the robustness of the identified potential biomarkers. Nevertheless, this work is a significant starting point in the search for safer, more accessible analytical tools for diagnosing prostate cancer.

Acknowledgements

FAPEMIG, CNPq and CAPES.

References

1. R.L. Siegel; A.N. Giaquinto; A. Jemal, *A Cancer Journal for Clinicians*. **2024**, 74, 17–48.
2. NATIONAL CANCER INSTITUTE. Prostate-Specific Antigen (PSA) Test. Bethesda, 2023. Available at: https://www.cancer.gov/types/prostate/psa-fact-sheet. Accessed on: June 24, 2025.

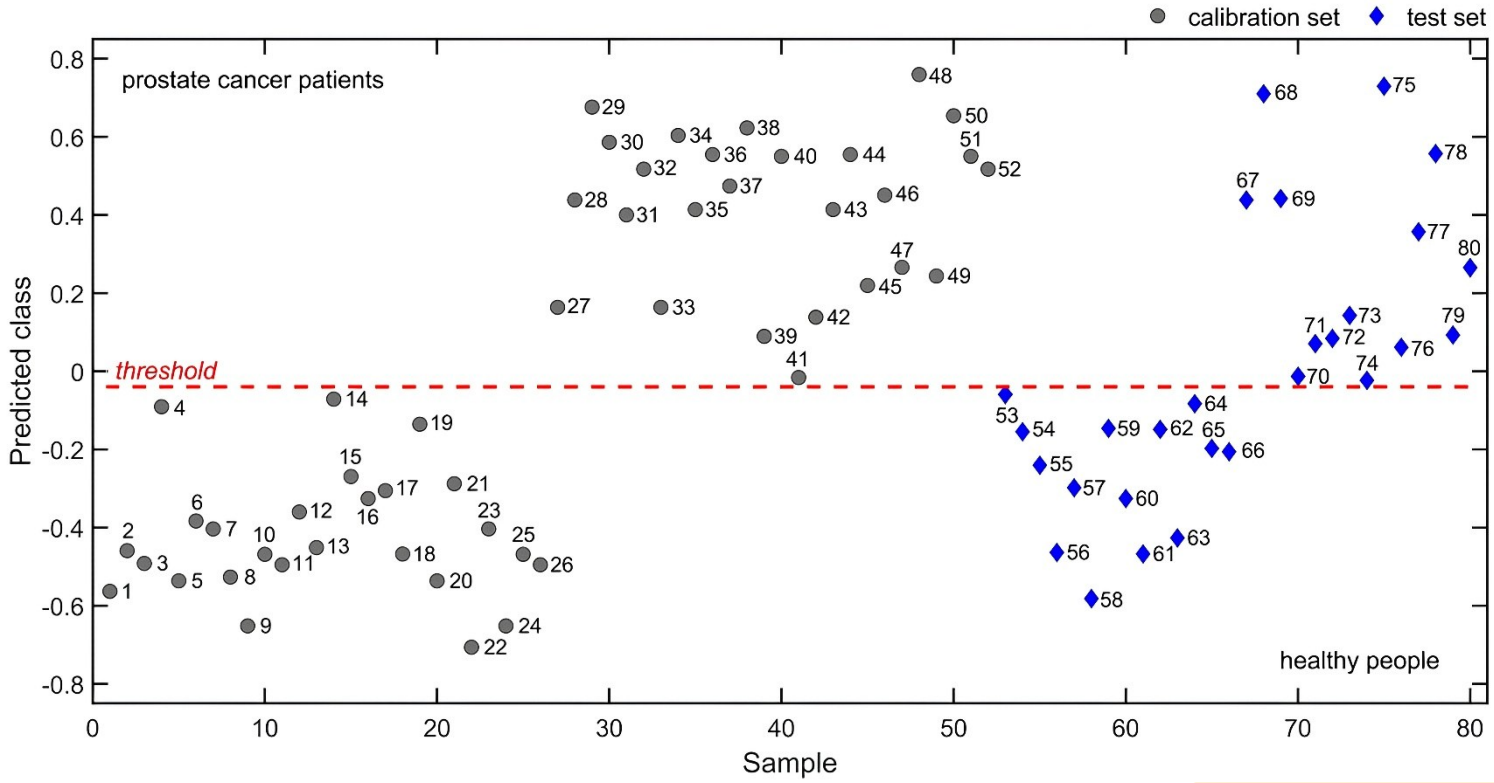

Figure 1. Figure 1 shows the class separation graph of the developed PLS-DA model. Samples above the dashed line belong to prostate cancer patients, while samples below the line belong to healthy individuals.