SDS
Statistica e
Data Science

Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia – Leonardo Egidi - Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

Questo volume è rilasciato sotto licenza Creative Commons
Attribuzione - Non commerciale - Non opere derivate 4.0

© 2024 The Authors

Contents

1 Keynote Sessions 11

1.1 Classification with imbalanced data and the (eternal?) struggle between statistics and data science. *Nicola Torelli* 11

1.2 Deep residual networks and differential equations. *Gérard Biau* 13

2 Invited - Complex data: new methodologies and applications . . 15

2.1 Link selection in binary regression models with the Model Confidence Set. *Michele La Rocca and Marcella Niglio and Marialuisa Restaino* 15

2.2 A cluster-weighted model for COVID-19 hospital admissions. *Daniele Spinelli, Paolo Berta, Salvatore Ingrassia and Giorgio Vittadini* 23

2.3 Multi-class text classification of news data. *Maurizio Romano and Maria Paola Priola* 28

3 Invited - Data science and dataspaces: challenges, results, and next steps 35

3.1 Data-Centric AI : A new Frontier emerging in Data Science. *Donato Malerba, Vincenzo Pasquadibisceglie, Vito Recchia and Annalisa Appice* 35

3.2 Data Spaces strategy to unleash agriculture data value: a concrete use case. *Nicola Masi, Delia Milazzo, Giulia Antonucci and Susanna Bonura*42

3.3 Addressing Agricultural Data Management Challenges with the Enhanced TRUE Connector. *Sergio Comella, Delia Milazzo, Mattia Giuseppe Marzano, Giulia Antonucci, Susanna Bonura and Angelo Marguglio* 48

4 Solicited - Data Science for Official Statistics 55

4.1 Data science at Istat for urban green. *Fabrizio De Fausti, Marco Di Zio, Giuseppe Lancioni, Stefano Mugnoli, Alberto Sabbi and Francesco Sisti* 55

4.2 Twitter (X) as a Data Source for Official Statistics: Monitoring Italian Debate on Immigration through Text Analysis. *Elena Catanese, Gerarda Grippo, Francesco Ortame and Maria Clelia Romano* 62

5 Solicited - Sustainable Artificial Intelligence in Finance 69

5.1 Feature Dependence and Prediction Explanations in P2P Lending. *Paolo Pagnottoni and Thanh Thuy Do* 69

6 Solicited - Young SIS 77

6.1 Merging data and historical information via optimal power prior selection in Bayesian models. *Roberto Macrì Demartino, Leonardo Egidi, Nicola Torelli and Ioannis Ntzoufras* 77

6.2 Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables. *Dalila Failli, Bruno Arpino, and Maria Francesca Marino* 84

6.3 A Simultaneous Spectral Clustering for Three-Way Data. *Cinzia Di Nuzzo and Salvatore Ingrassia* 90

7 Solicited - From Data Analysis to Data Science 97

7.1 Optimal Scaling: New Insights Into an Old Problem. *Gilbert Saporta* 97

8 Solicited - Statistical methods for textual data 101

8.1 PROCSIMA: Probability Distribution Clustering Using Similarity Matrix Analysis. *Marco Ortu* 101

8.2 Exploring Anti-Migrant Rhetoric on Italian Social Media. *Lara Fontanella, Annalina Sarra, Emiliano del Gobbo, Alex Cucco and Sara Fontanella* 108

8.3 Causal inference from texts: a random-forest approach. *Chiara Di Maria, Alessandro Albano, Mariangela Sciandra and Antonella Plaia* 114

9 Solicited - Data analysis methods for data in non-Euclidean spaces 121

9.1 Riemannian Statistics for Any Type of Data. *Oldemar Rodrıguez Rojas* 121

9.2 PAM clustering algorithm for ATR-FTIR spectral data selection: an application to multiple sclerosis. *Francesca Condino, Maria Caterina Crocco and Rita Guzzi* 128

9.3 Random Survival Forest for Censored Functional Data. *Giuseppe Loffredo, Elvira Romano and Fabrizio Maturo* 134

9.4 Advancing credit card fraud detection with innovative class partitioning and feature selection technique. *Mohammed Sabri, Antonio Balzanella and Rosanna Verde* 140

10 Solicited - Functional Data Analysis in Action 147

10.1 Functional Linear Discriminant Analysis for Misaligned Surfaces. *Tomas Masak* 147

10.2 Leveraging weighted functional data analysis to estimate earthquake-induced ground motion. *Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio* 155

10.3 Functional autoregressive processes on a spherical domain for global aircraft-based atmospheric measurements. *Almond Stöcker and Alessia Caponera* 161

11 Solicited - Bayesian Inference for Graphical Models 169

11.1 Log-likelihood approximation in Stochastic EM for Multilevel Latent Class Models. *Silvia Columbu, Nicola Piras and Jeroen K. Vermunt* 169

11.2 MCMC Sampling in Bayesian Gaussian Structure Learning. *Antonino Abbruzzo, Nicola Argentino, Reza Mohammadi, Maria De Iorio, Willem van den Boom and Alexandros Beskos* 176

12 Contributed - Promoting Equity: Statistical Insights into Tourism, Sustainability and Digital Divide 183

12.1 Lesson Learnt in the Data Science Worldview: New Dimension of Digital Divide. *Rita Lima* 183

12.2 An overview of Tourism Statistical Literacy. *Yasir Jehan, Giuseppina Lo Mascolo and Stefano De Cantis* 192

12.3 Scalable bootstrap inference via averaged Robbins-Monro approximations. *Giuseppe Alfonzetti and Ruggero Bellio* 198

12.4 The impact of sustainability on Initial Coin Offering: advantages in trading. *Alessandro Bitetto and Paola Cerchiello* 204

13 Contributed - High dimensional and functional data 211

13.1 Analysis of Brain-Body Physiological Rhythm Using Functional Graphical Models. *Rita Fici, Luigi Augugliaro and Ernst C. Wit* 211

13.2 A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects. *Ruggero Bellio and Cristiano Varin* 218

13.3 Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models. *Antonino Gagliano, Gianluca Sottile, Nicolina Sciaraffa, Claudia Coronnello and Luigi Augugliaro* 224

13.4 Investigating the association between high school outcomes and university enrollment choice: a machine learning approach. *Andrea Priulla, Alessandro Albano, Nicoletta D'Angelo and Massimo Attanasio* 230

14 Contributed - Statistical Analysis in economic and market dynamics 237

14.1 A comparison of multi-factor stochastic models for commodity prices C3. *Luca Vincenzo Ballestra, Christian Tezza and Paolo Foschi* 237

14.2 Nonparametric ranking estimation with application to the propensity for Circular Economy of Italian economic sectors. *Stefano Bonnini, Michela Borghesi and Massimiliano Giacalone* 246

14.3 Impact of the Russian invasion of Ukraine on coal markets: Evidence from an event-study approach. *Yana Kostiuk, Paola Cerchiello and Arianna Agosto* 252

14.4 Labour market and time series: a forecast approach for European countries from 1995 to 2022. *Paolo Mariani, Andrea Marletta and Piero Quatto* 258

15 Contributed - Innovations in cluster and latent class models ... 263
15.1 Biclustering of discrete data by extended finite mixtures of latent trait models. *Dalila Failli, Maria Francesca Marino and Francesca Martella* 263
15.2 Seismic events classification through latent class regression models for point processes. *Giada Lo Galbo, Giada Adelfio and Marcello Chiodi* 270
15.3 Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization. *Agostino Stavolo, Maria Gabriella Grassia, Marina Marino and Rocco Mazza* 276
16 Contributed - Modelling on spatial phenomena ... 283
16.1 Integrating computational and statistical algorithms in RT-GSCS for spatial survey administration. *Yuri Calleo, Simone Di Zio and Francesco Pilla* 283
16.2 Sensitivity mapping as a tool to support siting of offshore wind farms and increase citizens' acceptability. *Giovanna Cilluffo, Gianluca Sottile, Laura Ciriminna, Geraldina Signa, Agostino Tomasello and Salvatrice Vizzini* 290
16.3 Investigating hotel consumers' purchase intention on web analytics data through PLS-SEM. *Giuseppina Lo Mascolo, Chiara di Maria, Marcello Chiodi and Arabella Mocciaro Li Destri* 296
16.4 Spatio-temporal analysis of lightning point process data in severe storms. *Nicoletta D'Angelo, Milind Sharma, Marco Tarantino and Giada Adelfio*302
17 Contributed - Statistical machine learning for predictive modelling 309
17.1 Application of statistical techniques to predict the effective temperature of young stars. *Marco Tarantino, Loredana Prisinzano and Giada Adelfio* 309
17.2 Topological Attention for Denoising Astronomical Images. *Riccardo Ceccaroni and Pierpaolo Brutti* 316
17.3 LSTM-based Battery Life Prediction in IoT Systems: a proof of concept. *Vanessa Verrina, Andrea Vennera and Annarita Renda* 322
17.4 Predictive modeling of drivers' brake reaction time through machine learning methods. *Alessandro Albano, Giuseppe Salvo and Salvatore Russotto* 328
18 Contributed - Ordinal and preference data analysis . 335
18.1 OSILA (Order Statistics In Large Arrays): an original algorithm for an efficient attainment of the order statistics. *Andrea Cerasa* 335
18.2 The Mallows model with respondents' covariates for the analysis of preference rankings. *Marta Crispino, Lucia Modugno and Cristina Mollica* 343
18.3 Value-Based Predictors of Voting Intentions: An Empirical and Explainable approach. *Luca Pennella and Amin Gino Fabbrucci Barbagli* 349
18.4 A dynamic version of the Massey's rating system with an application in basketball. *Paolo Vidoni and Enrico Bozzo* 355
19 Contributed - Advances in text mining ... 361
19.1 Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?. *Andrea Sciandra and Arjuna Tuzzi* 361

19.2 A Fuzzy Topic Modeling approach to legal corpora. *Antonio Calcagnì and Arjuna Tuzzi* 368

19.3 EmurStat: a digital tool for statistical analysis of emur flow. *Simone Paesano, Maria Gabriella Grassia, Marina Marino, Dario Sacco and Rocco Mazza* 374

19.4 Graph Neural Networks for clustering medical documents. *Vittorio Torri and Francesca Ieva* 380

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data. Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science. In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized its 3rd international conference held in Palermo on the 11st and 12nd of April 2024, attended by over 100 researchers from different scientific fields. A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Palermo, April 11st and 12th, 2023

Antonella Plaia - Leonardo Egidi - Antonino Abbruzzo
Editors

EmurStat: a digital tool for statistical analysis of emur flow

Simone Paesano, Maria Gabriella Grassia, Marina Marino, Dario Sacco, Rocco Mazza

Abstract New Public Management (NPM) emphasizes the use of market-based techniques to improve efficiency and effectiveness in public service delivery. This approach seeks to promote accountability and performance measurement. Key performance indicators describe the performance of processes that characterize a specific workflow. One of the concepts that has emerged in the last decade is Precision Public Health, which integrates traditional determinants of health with new approaches such as data science and health economics. Moreover, visualizations help to understand social determinants of health and public health indicators. This paper aims to present a useful application for data visualization, processing, and analysis for understanding and evaluating the performance of services provided by emergency rooms, through the lunge on a specific case.

Key words: Data visualization, Emergency room.

Simone Paesano
University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: simone.paesano@unina.it

Maria Gabriella Grassia
University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: mgrassia@unina.it

Marina Marino
University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: marina.marino@unina.it

Dario Sacco
University of Naples Federico II, Department of Social Sciences, Vico Monte della Petà, 1, Napoli
e-mail: davide.sacco@unina.it

Rocco Mazza
University of Bari Aldo Moro, Department of Political Science, Piazza Cesare Battisi, Bari e-mail: rocco.mazza@uniba.it

1 Introduction

The New Public Management (NPM) represents a managerial approach applied in government agencies at both local and national levels to enhance the efficiency and effectiveness of public service organizations. Essentially, NPM entails the incorporation of principles and practices from the private sector into the public sector (Gomes & Mendes, 2022). Pollitt's taxonomy (1997) delineates eight interrelated components that encapsulate the objectives of NPM: cost reduction, budget constraints, transparent resource allocation, organizational disaggregation into separate agencies, decentralization of management authority within public agencies, separation of service provision from procurement, introduction of market-like mechanisms, and implementation of performance management incorporating goals, indicators, and results. NPM also advocates a shift in public employment practices, moving from tenure-based, nationally standardized pay and conditions to group contracts, performance-related pay, and local determination of pay and working conditions. The implementation of New Public Management (NPM) in healthcare has led to a significant emphasis on digitization processes and measurement of Key Performance Indicators (KPIs) (McGee-Lennon et al., 2017). There has been a growing momentum in digital health initiatives, with a focus on integrating digital technologies to improve health care delivery and improve patient outcomes (McGee-Lennon et al., 2017). Longitudinal qualitative evaluations have emphasized the importance of normalization process theory (NPT) and a comprehensive approach to framework analysis in assessing preparedness to deliver digital health care at scale (McGee-Lennon et al., 2017). In addition, the governance of digital transformation in health services has been recognized as a critical issue that requires in-depth evaluation and monitoring (Barros et al., 2019). In addition, the role of KPIs in digital health has been highlighted, with a call for the development of digital public health interventions and an assessment of the current state of knowledge in digital health from a management perspective (Wienert et al., 2022). This reflects a growing recognition of the need for robust measurement tools and frameworks to assess the impact and effectiveness of digital health initiatives. In addition, ethical issues related to digital health, such as the development of digital twins for personalized health services, have emerged, underscoring the complexities and considerations involved in the digitization of health services (Huang et al., 2022). In conclusion, the application of NPM in healthcare, particularly in the context of digitization processes and the importance of KPIs, is a multifaceted and evolving area. It requires a comprehensive understanding of preparedness, governance, ethical considerations, and the development of measurement tools to ensure the proper implementation and maintenance of digital health initiatives.

2 The guidelines on which the application is based.

The guidelines described by the Italian Ministry of Health can be considered as benchmarks for medical conduct (Mendogni et al., 2019). These guidelines have been instrumental in shaping various aspects of healthcare practices, including rehabilitation activities, pain management, clinical practice, and public health interventions. The Italian Ministry of Health has been actively involved in supporting projects, issuing national recommendations, and developing safety protocols based on these guidelines (Vitale et al., 2022). Specifically, this application was developed on three guidelines produced by the Ministry of Health, in the social health field with a focus on emergency room management. These are the national guidelines on intra-hospital triage; national guidelines on short intensive observation - obi and national guidelines on the development of the overcrowding management plan in emergency rooms. In addition, a session dedicated exclusively to socio-demographic analysis was produced. Before proceeding, it is beneficial to provide a brief definition of the focus of these guidelines. Triage is the initial point of reception for individuals arriving at the Emergency Department. The purpose of Triage is to oversee the management of all individuals accessing the Emergency Department, with particular attention to those in critical conditions. This involves assessing data and clinical indicators that characterize situations of risk, complications, and/or undesirable effects of ongoing treatments. This evaluation is conducted through clinical observation, the utilization of suitable scales/instruments, and appropriate monitoring systems. Triage services must be consistently available on a 24-hour basis.

2.1 Triage

Triage, is the initial point of reception for individuals arriving at the Emergency Department. The purpose of Triage is to oversee the management of all individuals accessing the Emergency Department, with particular attention to those in critical conditions. This involves assessing data and clinical indicators that characterize situations of risk, complications, and/or undesirable effects of ongoing treatments. This evaluation is conducted through clinical observation, the utilization of suitable scales/instruments, and appropriate monitoring systems. Triage services must be consistently available on a 24-hour basis. From a methodological standpoint, the Triage process unfolds in four phases: 1) Immediate Assessment Phase (at the entrance): This involves a swift observation of the individual's general appearance to identify those with care needs requiring immediate intervention. 2) Subjective and Objective Assessment Phase: a. Subjective assessment: Conducted through targeted interviews (focused medical history). b. Objective assessment: Carried out through the observation of clinical signs, vital parameter measurements, and analysis of available clinical documentation. 3) Triage Decision Phase: This entails assigning a priority code, implementing necessary care measures, and potentially activating diagnostic and therapeutic pathways. 4) Reassessment Phase: This phase involves

confirming or modifying the assigned priority code for patients in waiting, ensuring ongoing accuracy in the prioritization of cases. The statistical analyses included within the application are based on these methodological principles. The dimensions considered by the application are as follows: Mode of arrival; Absolute frequencies triage Igresso; Appropriateness triage code; Appropriateness intake times by triage code; Average waiting time and standard deviation grouped by code; Appropriateness by emergency department permanent times; Absolute frequencies type of trauma; Diagnosis code; Outcome pathway in emergency department (including patients in OBI).

2.2 Obi

Observation and Brief Intervention (OBI) is a management modality for emergency patients presenting with acute clinical conditions that are either high critical but low developmental risk, or low critical but with potential developmental risk. These cases have a high probability of reversibility and require an urgent and nondeferrable diagnostic and therapeutic process that cannot be managed in alternative healthcare settings. OBI is distinguished by its high intensity of care, involving significant medical and nursing staff commitment, diagnostic investigations, clinical monitoring, and strategic therapeutic planning, and is administered within a specifically defined and restricted time frame. Its goal is to ascertain the most appropriate level of care. The functions of the OBI include: - Clinical observation - Provision of short-term therapy for moderately complex conditions - Facilitation of diagnostic and therapeutic insights aimed at appropriate admission or discharge. The expected consequences are improved appropriateness of hospital admissions and safer discharge of patients from the emergency department. The three dimensions considered by the application are as follows: (i) Need for route in OBI. (ii) Appropriateness of OBI length of stay. (iii) Average time of admission to OBI and Standard deviation.

2.3 Overcrowding management plan in emergency rooms

Regarding these guidelines, the focus is on the concept of overcrowding, which serves as a reference for the other two discussed guidelines. Specifically, overcrowding is the situation where the normal functioning of Emergency Departments is constrained by the imbalance between healthcare demand (number of patients waiting and in care) and available logistical, instrumental, and professional resources. When overcrowding is persistent, it is generally due to the under-sizing of Emergency Departments about actual access, stemming from a real lack of structural, technological, human, and/or organizational resources. In this direction, the application shows a general data section where the total number of patients present within the emergency department in a single day is calculated.

3 Application workflow

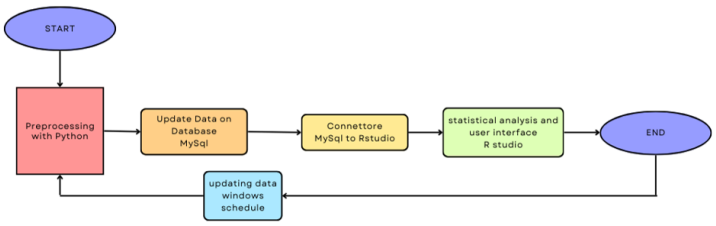

Fig. 1 Work flow apllication

Data mining, processing, and analysis play a key role in understanding and evaluating the performance of health services. These actions help discern patterns and trends and provide support for clinical and managerial decision-making (Vitale et al., 2022). This technical exposition outlines the complexities of the process of extracting and transforming health data. The main programming languages used are Python and R, while the database on which the application relies is MySQL. We used the Python programming language to facilitate the implementation of the data extraction process. This language boasts specialized libraries for adept data acquisition and processing, enabling access to diverse data sources such as CSV files, APIs, or databases. In healthcare, data extraction sources encompass electronic records, hospital management systems, and other health information systems. The specific case involves the extraction of files from Elipse based on The Urgent Emergency Information Flow (also called, Emur Flow), formatted as table format. After extraction, the data undergoes a series of transformations through Python. These encompass data cleaning, aggregation, and the generation of new variables. These operations are crucial for priming the data for subsequent analyses, ensuring consistency and quality. Following transformations, the data is securely stored in the MySQL database. Renowned for its scalability and performance, MySQL serves as a relational database management system, facilitating structured data retention and facile access for subsequent analyses. More intricate transformations, such as statistical analyses, are executed using the R programming language. R, widely acknowledged in statistics and data analysis, offers an array of specialized packages and functionalities. A monthly recall function, developed using Windows Scheduler, ensures the automatic loading of data into the MySQL Database.

4 Conclusion

The extraction and processing of healthcare data through Python, R, and MySQL yield indispensable insights for healthcare management and performance monitor-

ing. These tools enable the analysis of diverse healthcare datasets, contributing to a comprehensive understanding of healthcare systems and patient pathways. Moreover, the incorporation of standardized guidelines, particularly those from the Ministry of Health, safeguards the precision and consistency of statistical analyses (Mulugeta et al., 2019). By adhering to established guidelines, healthcare data analysts ensure the reliability and validity of their findings, thereby enhancing the quality of healthcare data analysis. Furthermore, the development of interactive dashboards facilitates effective data visualization, streamlining the understanding and interpretation of results (Radbron et al., 2021). Interactive dashboards provide a user-friendly interface for healthcare professionals and decision-makers to explore and comprehend complex healthcare data, thereby supporting informed decision-making and strategic planning. This integrative approach advances comprehension of the healthcare system, identifies avenues for enhancement, and buttresses data-driven decisions to uphold the quality and efficacy of healthcare services (Coates Mickan, 2020). In conclusion, the utilization of Data science tools for healthcare data analysis, in conjunction with adherence to standardized guidelines and the development of interactive dashboards, represents a comprehensive and robust approach to advancing healthcare management and performance monitoring.

References

1. Bourke, J. G. (2023). Scatalogic rites of all nations. BoD–Books on Demand.
2. Coates, D., Mickan, S. (2020). Challenges and enablers of the embedded researcher model. Journal of health organization and management, 34(7), 743-764.
3. Gomes, P., Mendes, S. M. (2023). Organizational context, use of performance management practices and their effects on organizational performance: an empirical look at these interrelationships. International Journal of Productivity and Performance Management, 72(8), 2467-2495.
4. Mulugeta, H., Wagnew, F., Dessie, G., Biresaw, H., Habtewold, T. D. (2019). Patient satisfaction with nursing care in Ethiopia: a systematic review and meta-analysis. BMC Nursing, 18, 1-12.
5. Pimentel-Nunes, P., Libânio, D., Marcos-Pinto, R., Areia, M., Leja, M., Esposito, G., ... Dinis-Ribeiro, M. (2019). Management of epithelial precancerous conditions and lesions in the stomach (maps II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy, 51(04), 365-388.
6. Pollitt, C., Dan, S. (2011). The impacts of the New Public Management in Europe: A meta-analysis.
7. Radbron, E., McCance, T., Middleton, R., Wilson, V. (2021). Maintaining momentum in action research. Nurse researcher, 29(3).
8. Tosi, D., Nosotti, M., Bonitta, G., Mazzucco, A., Righi, I., Mendogni, P., ... Crisci, R. (2019). Uniportal and three-portal video-assisted thoracic surgery lobectomy: analysis of the Italian video-assisted thoracic surgery group database. Interactive cardiovascular and thoracic surgery, 29(5), 714-721.
9. Visco, V., Vitale, C., Rispoli, A., Izzo, C., Virtuoso, N., Ferruzzi, G. J., ... Ciccarelli, M. (2022). Post-COVID-19 syndrome: involvement and interactions between respiratory, cardiovascular, and nervous systems. Journal of Clinical Medicine, 11(3), 524.