frontiers | Frontiers in Public Health

TYPE Perspective
PUBLISHED 22 March 2024
DOI 10.3389/fpubh.2024.1362699



OPEN ACCESS

EDITED BY
Ileana Baldi,
University of Padua, Italy

REVIEWED BY
Mario Gaio,
University of Campania Luigi Vanvitelli, Italy
Alexander Hochdorn,
University of Brasilia, Brazil
Rosaria Lombardo,
University of Campania Luigi Vanvitelli, Italy

*CORRESPONDENCE
Peter Kokol
peter.koko@um.si

RECEIVED 28 December 2023
ACCEPTED 07 March 2024
PUBLISHED 22 March 2024

CITATION
Žlahtič B, Kokol P, Blažun Vošner H and Završnik J (2024) The role of correspondence analysis in medical research.
*Front. Public Health* 12:1362699.
doi: 10.3389/fpubh.2024.1362699



# The role of correspondence analysis in medical research

Bojan Žlahtič[1], Peter Kokol[1,2]*, Helena Blažun Vošner[2,3] and Jernej Završnik[2,4]

[1]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia, [2]Community Healthcare Center dr. Adolf Drolc, Maribor, Slovenia, [3]Faculty of Health and Social Sciences Slovenj Gradec, Slovenj Gradec, Slovenia, [4]Alma Mater Europaea, Maribor, Slovenia

Correspondence analysis (CA) is a multivariate statistical and visualization technique. CA is extremely useful in analyzing either two- or multi-way contingency tables, representing some degree of correspondence between columns and rows. The CA results are visualized in easy-to-interpret "bi–plots," where the proximity of items (values of categorical variables) represents the degree of association between presented items. In other words, items positioned near each other are more associated than those located farther away. Each bi-plot has two dimensions, named during the analysis. The naming of dimensions adds a qualitative aspect to the analysis. Correspondence analysis may support medical professionals in finding answers to many important questions related to health, wellbeing, quality of life, and similar topics in a simpler but more informal way than by using more complex statistical or machine learning approaches. In that way, it can be used for dimension reduction and data simplification, clustering, classification, feature selection, knowledge extraction, visualization of adverse effects, or pattern detection.

KEYWORDS

**public health, medical research, correspondence analysis, exploratory data analysis, bibliometrics**

## 1 Introduction

This "perspective article" aims to demonstrate the usefulness of correspondence analysis (CA) and inform the medical community about possible CA benefits in research and everyday practice settings. Correspondence analysis is a multivariate statistical and visualization technique. When a contingency table consists of two variables, we talk about a simple CA; however, if the analysis is extended to more than two categorical variables, it is called a multiple CA (MCA). The roots of correspondence analysis date back to 1935 when Herman Otto Hartly (born Hirschfeld) published his work on contingency tables (1). Based on Hartly's work, Benzecri developed CA's mathematical foundations during the 1960s in France (2). However, the method became popular outside France through the work of Michael Greenacre (2) and Leabrt and coworkers (3). Greenacre and coworkers also popularized the use of CA in medical research (4). However, some attempts to use CA in healthcare have been made since 1975, mainly by French authors (5–9).

Variants of CA and MCA are extremely useful in analyzing either two- or multi-way contingency tables, representing some degree of correspondence between two or more categorical variables. They translate deviations from the independence model in a contingency table into distances. Conceptually, they are similar to principal component analysis but apply to categorical rather than continuous variables. CA and MCA enable users to graphically display row and column categories and visually inspect their associations. There are several extensions of CA and MCA (10), such as constrained, aggregate, or canonical correspondence analysis (5).

The CA results are presented in so-called "bi–plots," where the proximity of items (values of categorical variables) represents the degree of association between presented items. In other words, items positioned near each other are more associated than those located farther away. Each bi-plot has two dimensions, named during the analysis. The naming of dimensions adds a qualitative aspect to the analysis (6). It is worth noting that besides bi-plots, other graphical outputs, such as dendrograms or similarity trees, are commonly used when CA is employed for quali-quantitative analysis (7).

## 2 General benefits of correspondence analysis

The main benefits of CA are as follows:

- It shows relations and their strength between categorical categories in a way anyone can easily understand.
- It is objective because there are no underlying statistical distributional assumptions.
- It can be used on all types of categorical variables.
- It is a multivariate method.
- It provides a simple visualization of data.

## 3 Correspondence analysis in medical research and practice

To analyze the scope of CA use in medicine, we retrieved the corpus of publication from the Scopus indexing database (Elsevier, Netherlands). Scopus was chosen because it is considered reliable and authoritative and is the largest abstract and citation database of the research literature, including almost 50,000 source titles from more than 12,500 publishers. Scopus also covers MEDLINE and EMBASE databases and most of the Web of Science content. In addition, it provides advanced analytics services and enables 20,000 records to be exported simultaneously. The search was performed applying the following search string: {Correspondence analysis} in publication titles, abstracts, and keywords. The search was limited to the subject area of *Medicine*. The use of Curly Brackets {} denotes an exact search. In that way, we harvested 1,939 publications used in further bibliometric-based analysis (8).

The number of publications increased from 9 published in the year 1990 to 20 published in the year 2000, 63 in the year 2010, and 156 in the year 2022. The corpus of publications was analyzed using synthetic knowledge synthesis, a triangulation of bibliometrics, bibliometric mapping, and content analysis (11). For this perspective study, bibliometric mapping was performed using VOSViewer software, version 1.6.20 (9); however, other mapping software, such as Bibliometrix (12), exist, which could be used similarly. Bibliometric mapping on authors' keywords resulted in four clusters, represented by different colors, as shown in Figure 1. By applying content analysis to cluster terms, we identified four themes, each presenting the use of correspondence analysis in Medicine. The themes, together with influential and interesting studies, are explained below:

- *Correspondence analysis in genetics (light blue cluster):* Correspondence analysis was used to investigate the relationship between transcriptional programs of the osteoarthritis genetic landscape and clinical outcomes using the severity index (13). In a retrospective study regarding clinical pathological characteristics and outcomes of triple-negative breast cancers, correspondence analysis was used in the investigation of the relationship between androgen receptor protein expression, core-needle biopsy (using different cutoffs), and standard clinical and pathological variables, including stromal tumor-infiltrating lymphocytes (14).
- *Multiple correspondence analysis combined with machine learning (yellow cluster):* Multiple correspondence analysis and random forests were used to analyze the linkage among socio-demographic, behavioral, psycho-social, and biological factors

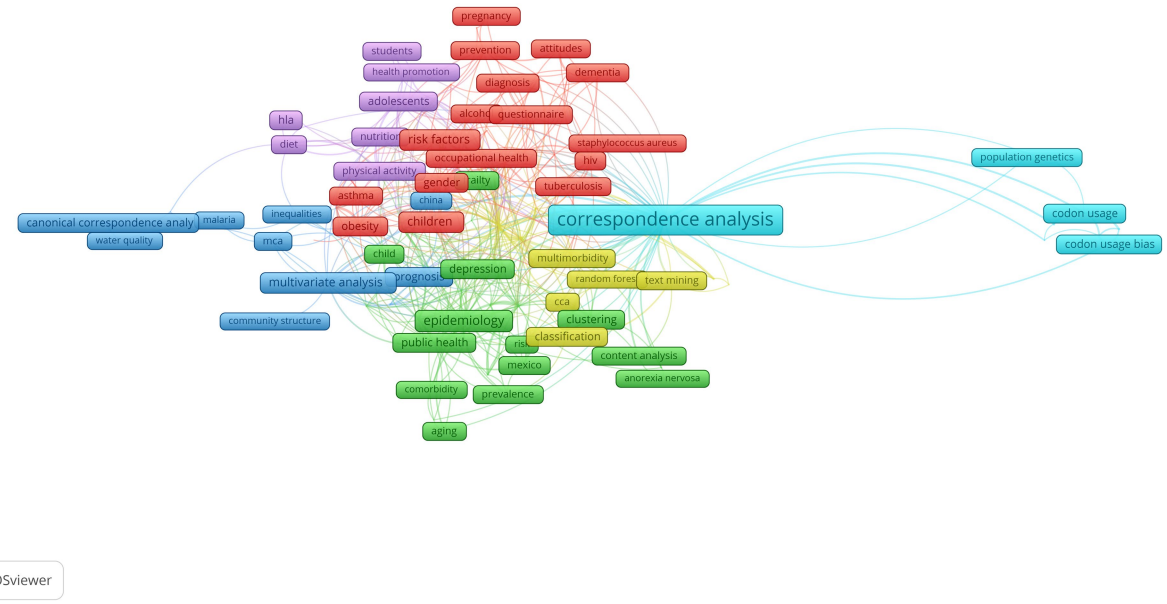


FIGURE 1
The research landscape of the role of correspondence analysis in medical research and practice.

associated with high HIV RNA viral load (15). Data from the first wave of COVID-19 in New Zealand were analyzed comprising PCR-confirmed and symptomatic PCR-negative individuals using multiple correspondence analysis in combination with various machine learning algorithms (11).

- *Epidemiology and public health (green and blue cluster):* Responses to the questionnaire regarding policies, guidelines, civil awareness, epidemiology and data, detection rate, and care management of 102 countries were analyzed using multiple correspondence analysis to asses the country preparedness for management of non-alcoholic fatty liver disease (16). An online survey administered to the Italian population was analyzed using multiple correspondence analysis to detect the factorial dimensions underpinning ways of interpreting the social environment regarding decision-making, people's mindsets, and similar aspects during the COVID-19 pandemic (17).
- *Healthy and active living (violet cluster):* Multiple correspondence analysis was used to classify youths into mental health profiles. Adolescents were categorized into three mental health profiles based on their mental wellbeing, resilience, quality of life, cognitive and behavioral disorder symptoms, and use of tobacco, alcohol, and similar substances (18). Multiple correspondence analysis combined with clustering was used to explore potential changes in dietary intake, physical activity, body weight, and food supply relative to individual characteristics during the COVID-19 lockdown (19).
- *Multiple correspondence analysis in primary healthcare (red cluster):* Multiple correspondence analysis was used to analyze the association of physical health multimorbidity in patients with and without severe mental illness (20). Multiple correspondence analysis was also applied to classify asthma into six subtypes based on data gathered from a large number of longitudinal primary care electronic health records (21).

## 4 Dissuasion: benefits of using correspondence analysis in medical research

In addition to the general benefits of using correspondence analysis addressed above, healthcare professionals with just a short training in CA can use it for complex tasks such as:

- dimension reduction and data simplification
- clustering
- classification
- feature selection
- knowledge extraction
- qualitative component
- visualization of adverse effect
- pattern detection

Figure 2 demonstrates the results of the use of multiple correspondence analysis on the STROKE database. The STROKE

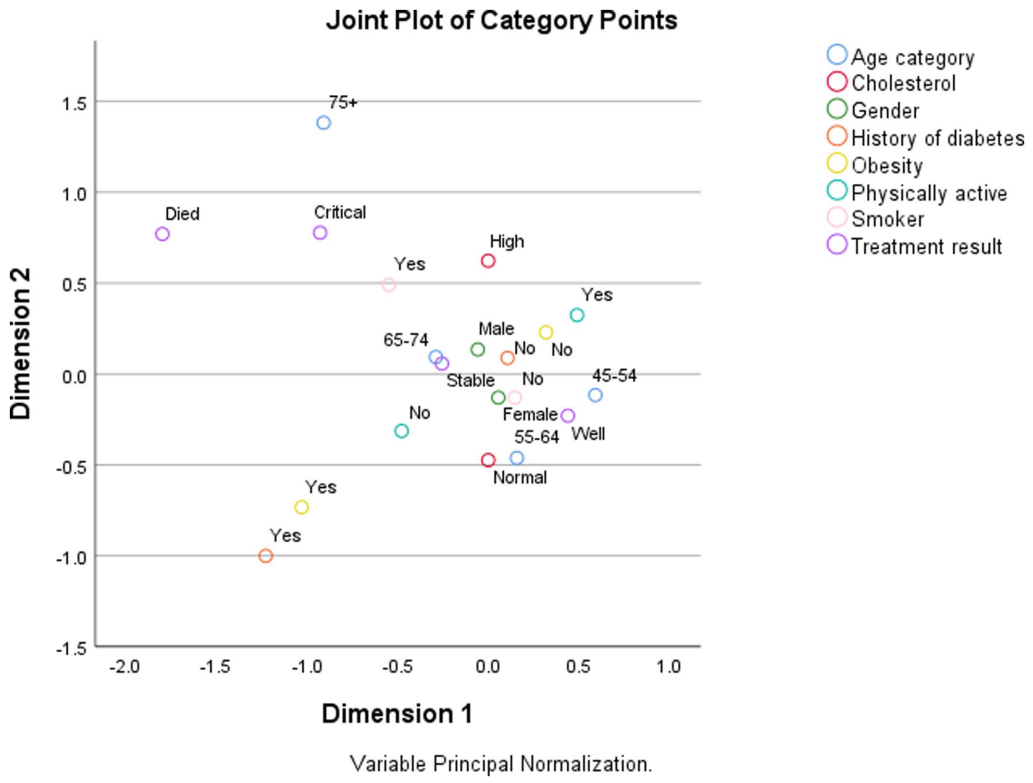


FIGURE 2
An example of using correspondence analysis in medical research.

database is provided by SPSS (IBM, Rochester, United States) in its sample set. It contains cleansed medical data for approximately 2,412 stroke patients collected from 20 hospitals. The data consist of demographic variables such as sex, age, physical activity, smoking status, health status variables including the presence of obesity, high cholesterol, and diabetes, and finally, the variable presenting the treatment outcome. The analysis was performed with SPSS, Version 29 (IBM, Rochester, United States). Figure 2 reveals that physically active women younger than 64 years with normal cholesterol are located near the treatment result "Well" category. On the other hand, patients older than 75 years and smoking are located near the treatment result "Critical" category. This evidence can be used for classification or clustering. The same evidence shows that age, smoking, cholesterol, and physical activity are important variables (Feature selection). On the other hand, the values of obesity and history of diabetes are located far from the treatment result in the "Critical or Death" category, and the male and female sexes are located very near each other, which might indicate that those variables are not so crucial for stroke management (dimension reduction and data simplification). The above evidence also contributes to pattern recognition and knowledge extraction. Dimension 1 could be labeled age and treatment results, which could qualitatively mean that the treatment results after stroke will worsen with aging. Dimension 2 could be labeled high cholesterol and smoking, qualitatively meaning that both have considerable adverse effects on stroke management.

## 5 Conclusion

The use of correspondence analysis in medicine is growing exponentially. As revealed in our study, it is employed in an increased number of different medical contexts for various tasks. Based on our findings, we believe that correspondence analysis may support medical professionals in finding answers to many important questions related to health, wellbeing, quality of life, and similar topics in a more straightforward but informal way than by using more complex statistical or machine learning approaches.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

BŽ: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. PK: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. HB: Supervision, Validation, Writing – original draft, Writing – review & editing. JZ: Supervision, Validation, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note



## References

1. Hirschfeld HO. A connection between correlation and contingency. *Math Proc Camb Philos Soc*. (1935) 31:520–4. doi: 10.1017/S0305004100013517

2. Greenacre MJ ed. *Theory and applications of correspondence analysis*. London; Orlando, FL: Academic Press (1984). 364 p.

3. Lebart L, Morineau A, Warwick KM. *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. Hoboken, NJ: Wiley (1984). 264 p.

4. Greenacre M. Correspondence analysis in medical research. *Stat Methods Med Res*. (1992) 1:97–117. doi: 10.1177/096228029200100106

5. Beh EJ, Lombardo R. *Correspondence analysis: Theory, practice and new strategies*. *1st* ed. Chichester, West Sussex; Hoboken, NJ: Wiley (2014). 592 p.

6. Kokol P, Blažun Vošner H, Železnik D. Visualising nursing data using correspondence analysis. *Nurse Res*. (2016) 24:38–40. doi: 10.7748/nr.2016.e1441

7. Canuto A, Braga B, Monteiro L, Melo R. Aspectos críticos do uso de caqdas na pesquisa qualitativa: Uma comparação empírica das ferramentas digitais alceste e iramuteq. *New Trends Qual Res*. (2020) 3:199–211. doi: 10.36367/ntqr.3.2020.199-211

8. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. *J Informet*. (2017) 11:959–75. doi: 10.1016/j.joi.2017.08.007

9. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*. (2010) 84:523–38. doi: 10.1007/s11192-009-0146-3

10. Blasius J. Correspondence analysis In: M Lovric, editor. *International encyclopedia of statistical science*. Berlin, Heidelberg: Springer (2011). 318–21.

11. French N, Jones G, Heuer C, Hope V, Jefferies S, Muellner P, et al. Creating symptom-based criteria for diagnostic testing: a case study based on a multivariate analysis of data collected during the first wave of the COVID-19 pandemic in New Zealand. *BMC Infect Dis*. (2021) 21:1119. doi: 10.1186/s12879-021-06810-4

12. Belfiore A, Cuccurullo C, Aria M. IoT in healthcare: a scientometric analysis. *Technol Forecast Soc Change*. (2022) 184:122001. doi: 10.1016/j.techfore.2022.122001

13. Ji Q, Zheng Y, Zhang G, Hu Y, Fan X, Hou Y, et al. Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. *Ann Rheum Dis*. (2019) 78:100–10. doi: 10.1136/annrheumdis-2017-212863

14. Jongen L, Floris G, Wildiers H, Claessens F, Richard F, Laenen A, et al. Tumor characteristics and outcome by androgen receptor expression in triple-negative breast cancer patients treated with neo-adjuvant chemotherapy. *Breast Cancer Res Treat*. (2019) 176:699–708. doi: 10.1007/s10549-019-05252-6

15. Soogun AO, Kharsany ABM, Zewotir T, North D, Ogunsakin RE. Identifying potential factors associated with high HIV viral load in KwaZulu-Natal, South Africa

using multiple correspondence analysis and random Forest analysis. *BMC Med Res Methodol*. (2022) 22:174. doi: 10.1186/s12874-022-01625-6

16. Lazarus JV, Mark HE, Villota-Rivas M, Palayew A, Carrieri P, Colombo M, et al. The global NAFLD policy review and preparedness index: are countries ready to address this silent public health challenge? *J Hepatol*. (2022) 76:771–80. doi: 10.1016/j.jhep.2021.10.025

17. Gennaro A, Reho M, Marinaci T, Cordella B, Castiglioni M, Caldiroli CL, et al. Social environment and attitudes toward COVID-19 anti-contagious measures: an explorative study from Italy. *Int J Environ Res Public Health*. (2023) 20:3621. doi: 10.3390/ijerph20043621

18. Las-Hayas C, Mateo-Abad M, Vergara I, Izco-Basurko I, González-Pinto A, Gabrielli S, et al. Relevance of well-being, resilience, and health-related quality of life to mental health profiles of European adolescents: results from a cross-sectional analysis of the school-based multinational UPRIGHT project. *Soc Psychiatry Psychiatr Epidemiol*. (2022) 57:279–91. doi: 10.1007/s00127-021-02156-z

19. Deschasaux-Tanguy M, Druesne-Pecollo N, Esseddik Y, De Edelenyi FS, Allès B, Andreeva VA, et al. Diet and physical activity during the coronavirus disease 2019 (COVID-19) lockdown (March-May 2020): results from the French NutriNet-Santé Cohort Study. *Am J Clin Nutr*. (2021) 113:924–38. doi: 10.1093/ajcn/nqaa336

20. Launders N, Hayes JF, Price G, Osborn DPJ. Clustering of physical health multimorbidity in people with severe mental illness: an accumulated prevalence analysis of United Kingdom primary care data. *PLoS Med*. (2022) 19:e1003976. doi: 10.1371/journal.pmed.1003976

21. Horne EMF, McLean S, Alsallakh MA, Davies GA, Price DB, Sheikh A, et al. Defining clinical subtypes of adult asthma using electronic health records: analysis of a large UK primary care database with external validation. *Int J Med Inform*. (2023) 170:104942. doi: 10.1016/j.ijmedinf.2022.104942