Acta Scientific NUTRITIONAL HEALTH (ISSN:2582-1423)

Volume 8 Issue 1 January 2024

Research Article

Statistical Analysis of Risk Factors with Overweight-Diabetes and Gut Microbiota, a Bibliometric Analysis Using Bibliometrix

Tell-Morinson Melba*, Menco-Tovar Andrea, Barrero-Jiménez Erika, Moreno-Novoa Melissa and Mendez-Ramos Maria

Statistical Research and Applied Mathematical Modeling Group (GEMMA), Department of Mathematics, Faculty of Education and Sciences, University of Sucre, Colombia

***Corresponding Author:** Tell-Morinson Melba, Statistical Research and Applied Mathematical Modeling Group (GEMMA), Department of Mathematics, Faculty of Education and Sciences, University of Sucre, Colombia.

DOI: 10.31080/ASNH.2023.08.1334

Received: October 28, 2023

Published: December 04, 2023

Abstract

Sceintific production on chronic non-communicable diseases such as overweight, obesity and diabetes has grown in recent years. However, there is no comprehensive overview of the study designs and statistical analysis methodologies commonly used for risk assessment of research conducted on this topic. In order to evaluate risk assessment in this field, 1190 relevant documents downloaded using markers were included: ("Risk-Factors Associated with Overweight" or "Risk-Factors Associated with Obesity" or "Risk-Factors Associated with diabetes") and (microbiome or diabetes) and "multivariate analysis", in publications retrieved between 2017-2022 from the PubMed database; Specific parameters of title, journal, year of publication, authors, country of origin, institution, authors, keywords, among others, were analyzed. Data analysis was performed in three stages: 1. Descriptive; 2. Networking for Bibliographic Coupling Analysis (Network); 3. Strength of association of bibliographic coupling (Normalization). Data Visualization comprised: a. A mapping of the conceptual structure with Multiple Correspondence Analysis (MCA) for qualitative variables; b Network mapping. Grouping (Clustering) of K-means to identify groups of documents that express common concepts. To automate the data analysis and visualization stages, the open source tool (bibliometrix R-package), developed in R language, was used. 68.6% of the variability of the information was explained in the first factorial plane ACM, the cluster (cluster) of K-means identified two groups of documents that express common concepts. Cross-sectional, case-control, cohort and clinical trials are presented. "Risk" defined as the probability of occurrence of a clinical event. Use of models such as: logistic regression to relate explanatory variables of the risk of the occurrence of the event over a period of time; or, Cox proportional hazards regression, to relate explanatory variables of the conditional instantaneous risk of the event; Diagnostic tests for the detection of the clinical event (sensitivity, specificity, predictive value and likelihood ratio) and the use of the ROC curve.

Keywords: Gut Microbiota; Overweight; Obesity; Diabetes; Multivariate Analysis

Introduction

Some low- or lower-middle-income countries have gender inequity and inequality for women, which means that the biological characteristics of the mother have a negative influence on the mental development (learning) and motor development of her children, where the microbiota will have a direct effect on the brain [1]. Parents greatly influence the behaviors their children adopt often do not recognize the risk of excessive weight gain in children and youth, including those that affect weight (e.g., diet and physical activity). Malnutrition due to excess (overweight or obesity) is an epidemic that is advancing day by day in the general population. Among the leading causes of death worldwide are obesity and diabetes. The development of new innovative intervention strategies from academia that have collaborative learning and ICTs as a transversal axis to prevent diseases such as obesity and diabetes in adults and children. Studies show biological, genetic, environmental factors, as well as the gut microbiota as agents involved and consequences such as diabetes (DOI: 10.1007/s12519-019-00267-).

In statistics, we are concerned with measuring characteristics (variables) of people in a population, understanding properties of the distribution of the variables studied (e.g., central tendency, variability), and studying the relationships between different variables. Most study methods – study design, analysis methodology – are based on continuous measurement variables, commonly using methods based on distributions such as normal for statistical analysis. In clinical epidemiology, where statistical methodology is applied to the study of diseases, their occurrence, distribution and relationship with explanatory variables, confronts us with variables of interest with distributions that are not continuous. Commonly continuous measurement variables become dichotomous

to facilitate their understanding, for example, a person's blood glucose level is a continuous measurement variable, but a cut-off point is set above which the individual is considered "glucose intolerant" and therefore requires treatment, while another person with glucose below the cut-off point, It does not require treatment. If some clinical variables become dichotomous in the field of application in epidemiology, this article will explore the statistical methods of study designs and analysis methodologies for this type of risk. Given the high degree of complexity of this problem and the mathematical foundation implicit in it, we are interested in analyzing from a *Data Science* perspective, as a trend evaluation tool, a valuable analytical technique to map existing literature on a specific research topic (scientific mapping). The *bibliometric* approach was chosen for qualitative-quantitative literature (systematic, transparent and reproducible review process based on the statistical measurement of science or scientific activity). The use of bibliometrics has spread to all disciplines (https://doi.org/10.1016/S0306-4573(98)00028-4), one of the areas including childhood obesity [2]. Considering high-impact publications, actively collect evidence from previous research papers, of connections between authors, frameworks, methodology, and practice of everything that is published on this topic [3].

This study was developed considering three levels of analysis: sources, authors, and documents in an objective and reliable manner. In the first instance, the study focused on identifying the relevance of the topics related to each level, understanding relevance as the most productive or cited item, depending on the unit of analysis. Second, knowledge structures were developed, which used various bibliometric techniques. Specifically, conceptual structures focused on major themes and trends, intellectual structures on how specific works influence a scientific community, and the social structure that shows collaboration between authors and countries. Co-occurrence networks, collaborative networks, thematic maps, and world maps are provided. A network is a graphical representation of the co-occurrences of elements in a set of documents. In a co-occurrence network, items consist of terms drawn from the article's keyword lists, titles, or abstracts; whereas, in a collaborative network, the elements consist of co-authors, author affiliations, or author countries. A thematic map is a Cartesian representation of the clusters of terms identified by performing cluster analysis in a co-occurrence network. It allows for easier interpretation of research topics developed in a framework. Finally, a world map is a geographical representation of the collaborative network of an author's country. The analyses were based on Keywords Plus, which are words or phrases that appear frequently in the titles of references cited in an article but do not appear in the title of the article itself. They are extracted from articles using a statistical algorithm, based on the references cited in the article.

Data analytics according to the literature is proposed in three stages: Descriptive; Networking for Network analysis, co-citation, collaboration and co-occurrence; Strength of association of bibliographic coupling, co-citation and co-occurrence data (Normalization). Data visualization comprises: a. a mapping of the conceptual structure (Statistical analysis of research on Risk factors associated with gut microbiome and metabolic disorders such as obesity and diabetes of the last 5 years) with weighted principal component analysis - Multiple Correspondence Analysis (MCA) for qualitative variables; b Network mapping. Grouping (Clustering) of K-means to identify groups of documents that express common concepts). The MCA multivariate technique seeks to reduce variability in the group of individuals [1] in order to identify more accurately and easily those with the greatest influence on the study [4].

This paper presents a bibliometric analysis of the empirical literature on the analysis of data from risk studies where gut microbiota, obesity and/or diabetes are related to multi-factors (internal and/or environmental); Study designs and statistical analysis methodologies commonly used for risk assessment. The result of the analysis deploys most frequently used keywords using word cloud and conceptual structure mapping method to provide the scientific research community with a comprehensive understanding of the effect of overnutrition globally. The specific objectives of the analysis include identifying the growth of scientific research, publications, and citation trends over time for statistical analysis of research on risk factors associated with gut microbiome and metabolic disorders such as obesity and diabetes over the past 5 years; Identify countries that contributed the most, active journals, authors, institutions that improve research in statistical analysis of research on risk factors associated with gut microbiome and metabolic disorders such as obesity and diabetes.

Methodology

Study design

Bibliometric analysis, a quantitative approach to analyzing academic literature using bibliographies, provides description, evaluation, and follow-up of published research from scientifically indexed documents, using it as a tool for research and information collection. The database chosen for this study was *PubMed*, specialized in health and open access, constantly updated with millions of bibliographic references, from more than 5,000 scientific journals and from different countries, allowing simple or complex searches based on search functions by fields. In addition, search strategies can be saved in different tools and formats, as well as creating alerts and archiving the results [5]. Using the power of the *PubMed* database, it allowed us to limit chronological periods, type of publication, publications with the highest impact, and other parameters. The documents considered were in *"Full text"* and "*Free full tex*" modalities with criteria for analysis of the production from 2017 to September 20, 2022. The methodological objective was to analyze publications, citations and sources of information to analyze trends in the statistical analysis of research on risk factors associated with gut microbiota and metabolic disorders such as obesity and diabetes in the last 5 years. Our findings are interpreted and described in the chapter on Analyzing the Results. The standard general scientific mapping workflow (Figure 1) consisted of the following stages: Study design; Data collection and analysis; Data Visualization and Interpretation.

The search equation: ("Risk-Factors Associated with Over-

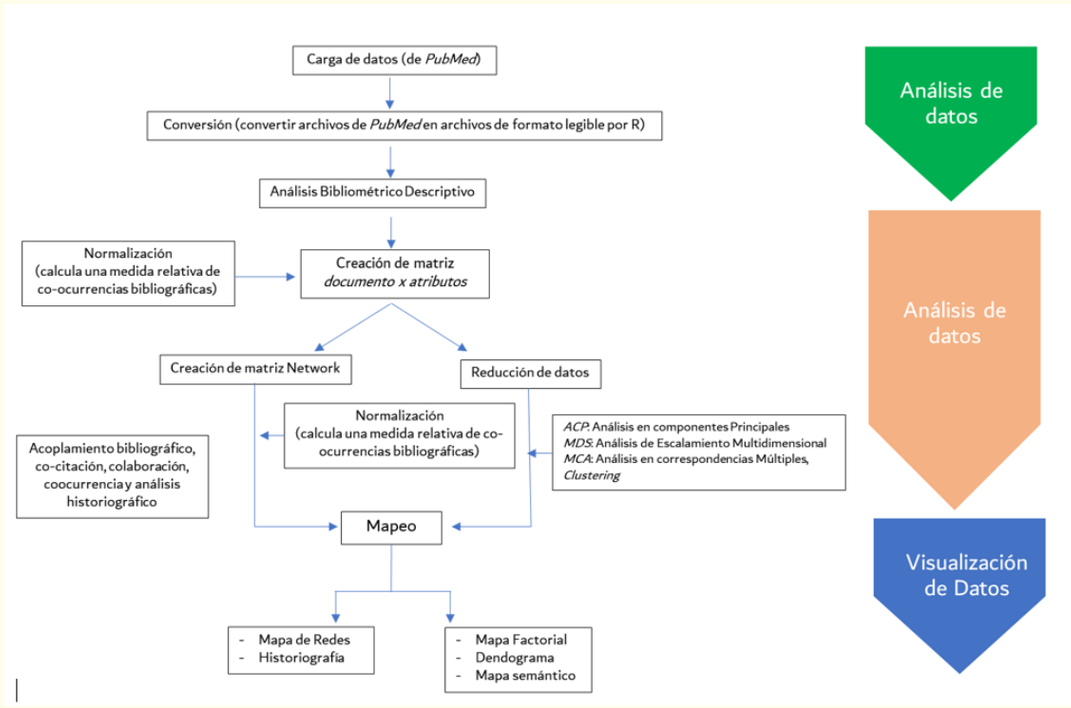


Figure 1: Scientific mapping workflow with Bibliometrix [6].

weight" or "Risk-Factors Associated with Obesity" or "Risk-Factors Associated with diabetes") and (microbiome or diabetes) and "multivariate analysis", according to the results obtained, the relevant publications in the chosen population, the most important scientific journals, publications by countries, publications by organizations and institutions, impact and relevance of authors (N = 1190) were taken. This search was conducted from 2017 to 20 September 2022. Although language was not used as a filter, it should be noted that the search was conducted using English, which could be understood as a "quasi-filter". However, in many cases, at least the title, abstract, and keyword are written, in addition to a specific language, in English. Therefore, language was not considered a limitation. From 2017 to September 20, 2022, scientific production in the field of statistical analysis with multivariate techniques, risk factors associated with gut microbiota, obesity, diabetes experienced significant fluctuations in terms of publication volume (Figure 2). In fact, although the average annual production was over 100, the production over the years was not uniform. In terms of production per year, the database was divided into four periods, ensuring a maximum production per year in the First Period below 120, a threshold that was exceeded in 2019. Consequently, the First Period (Period 1) was from 2017 to 2019, while the last Period (Period 4) was from 2021 to the present day, with a drop from 250 to 120 approx. It should also be noted that, in some results, the Total Period was used, corresponding to the entire period studied, from 2017 to 2022.

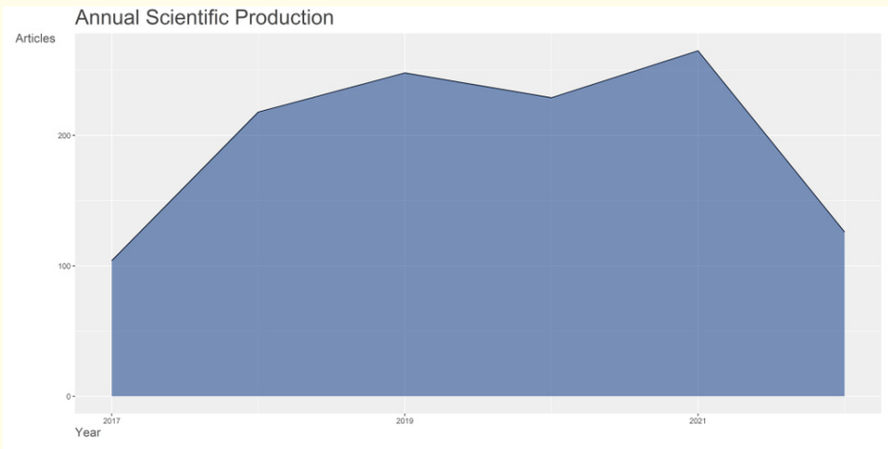


Figure 2: Annual scientific production for the study.

Among the 1190 articles, 50 articles were delimited for the construction of the results taking into account the following criteria: Where the related words previously placed for the placement of criteria of the article in PubMed are identified. Of the 50 articles, every 10 were separated in order to be able to better evaluate them, which with the evaluation after reading the abstracts showed only 16 articles that satisfy the answer to the problem question proposed in this study, for the article and as such called the systematic review. In the second sub-stage, in which some records were excluded, the information was leaked. The results of documents exported and analyzed from relevant studies were archived in ".txt" format, for further processing and analysis.

In the third sub-stage, the database was loaded and converted into a bibliometrix data framework, duplicate data was removed. Finally, in the fourth sub-stage, records obtained before 2017 were excluded as that year was the starting point of the analysis.

Data analysis

- **Stage 1:** *Descriptive analysis* of the bibliographic data framework: Articles, citations per article Authors, Author appearances, Authors of 1-author articles - multiple authors, Articles per author, Index of collaboration of co-authors per article, Top 10 - Most productive authors, Top 10 (Most cited articles, most productive countries -according to first author affiliation-, Most frequent journals, Most frequent keywords, Citation Analysis: Most Cited References and Top 10 Most Cited Authors).
- **Stage 2:** *Networking for Network analysis, co-citation, collaboration and co-occurrence*. The size of the node is proportional to the number of documents the country holds. If there is a connection between the nodes, it indicates that there is a cooperative relationship between the two countries. Collaborative scientific analysis was used to identify the social structure of the field, by applying social network analysis [7], applying it at the aggregate level (i.e., countries).
- **Stage 3:** Standardization. Taxonomy bibliometric techniques were used for bibliographic coupling (author, document and Journal), co-citation (author, reference and Journal) and co-occurrence data (author, country of affiliation and institution of affiliation), association (proximity index) was sought as a measure of similarity [8].

Data Visualization comprised

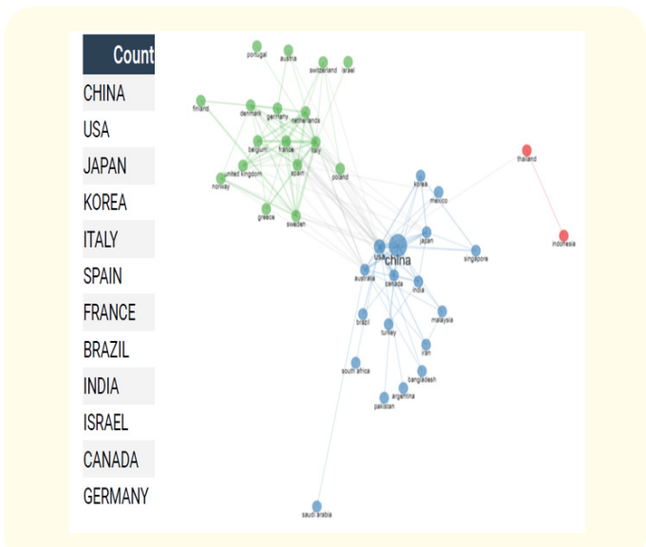
- Multidimensional scaling analysis of high-frequency keywords for the field of Statistical Analysis of Research on Risk Factors Associated with Overweight-Diabetes and Gut Microbiome, was the assignment of observational data to a specific location in conceptual space (generally, two-dimensional or three-dimensional).
- Mapping of the conceptual structure ("diabetes", "microbiota", "obesity" and "multivariate analysis"): principal component analysis weighted with Multiple Correspondence Analysis (MCA) for qualitative variables, to draw a structure of the conceptual field [9-11]); Grouping (Clustering) of K-means to identify groups of documents that express common concepts). In this paper, the statistical clustering method was used to calculate the co-word matrix, and keywords with a relatively high concurrency frequency were grouped into small groups, and keywords with a relatively low frequency were grouped into large groups. We form a tree dendogram of relationships from close to alienated.

The productivity indices (PI) of authors, citations of authors and articles, h-index, g-index or p-index were not recorded in this study due to the limitation of the *PubMed* database. To automate the data analysis and visualization stages, the open-source tool bibliometrix R-package (http://www.bibliometrix.org) was used, developed in language [6], a recent R package that facilitates a more complete bibliometric analysis using specific tools for both bibliometric and scientometric quantitative research [12,13].

Results and Analysis

To carry out this analysis, various bibliometric techniques explained in the methodology have been put into practice with the intention of analysing the significance at the research level. The documents that were considered for the study were scientific articles in "*Full text" and "*Free full text" mode, in order to segment the analysis only to articles in complete versions. The results presented here refer to the 1,190 documents analyzed after the bibliographic search: 700 of these were full *Journal* articles, the rest of the publications are associated with comparative, observational, multicenter studies, very few review studies, systematic reviews or Meta-Analyses that relate the gut microbiota, obesity and diabetes, associated with multivariate analysis.

Some 77 countries contributed to publications on the topic (Figure 3). China was the most productive country (NP = 288), followed by the United States (NP = 81), Japan (NP = 75), and Korea (NP = 64) were at the top, four influential countries (Figure IP). Meanwhile, China was the most productive country based on publications (Inter-Country Collaboration).



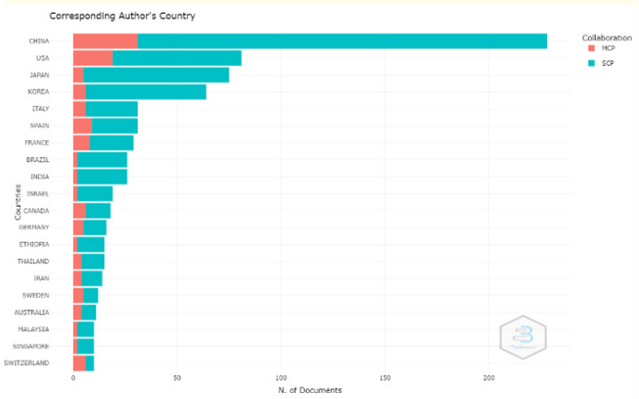


Figure 3: The relationship between countries for the subject of study and distribution of the Collaboration Index within (SCP) and between (CCM) countries with the highest scientific production. Source: Authors.
SCR: Standard Competition Ranking

The countries with the most referenced documents and whose contribution and impact is more outside (CCM) than inside them (SCP) are: China, the United States, Japan and Korea. The Top 10: most productive countries (according to the first author's affiliation) are: China, United States, Japan, Korea, Italy, Spain, France, Brazil, India, and Israel.

It was found that the country with the most collaborations was China. Its main networks are the United States and Australia. The second was Italy, which has scientific research in conjunction with Germany. These types of relationships allow us to show that the contribution to the subject is being to some extent driven by mechanisms for the generation of new knowledge from different territorial contexts and from an international approach (*Analysis of cooperation networks*).

The frequency of words in the 1,190 documents analyzed shows that the 10 *most relevant words (*Keyword plus, Title's Word and Abstract Word) by indicator are related to risk factors taking into account sex, retrospective studies, multivariate analysis related to the human population segmented by age groups (adolescent, youth, middle age, adult and older adult), thus setting the profile of research on the subject. A graphic example of the most repeated words in scientific texts can be found in the following word cloud (Figure 4): *Title's Words.* Patients (431 occurrences), Risk (352), study (318), factors (300), diabetes (267), disease (177), type (151), mellitus (122), association (102) and cohort (95); *Abstract´s Words.* Patients (4,613 occurrences), Risk (2,829), diabetes (2,092), factors (1,963), CI (1,884), study (1,871), analysis (1,754), age (1,457), multivariate (1,250) and disease (1,174); *KeyWords Plus.* Humans (943 occurrences), female (816), male (770), risk factors (721), middle aged (674), aged (556), adult (423), multivariate analysis (388), retrospective study (355) and aged 80 and over (205).

Figure 4: a) Frequency of most commonly used words in scientific texts on the gut microbiome in obesity and diabetes, associated with multivariate analysis; b) Network of words (thematic map). Zurce: Authors.

Thematic Map

Conceptual Structure (co-occurrence network): Cobo., *et al.* (2011) made a proposal to measure a specific field of research conceptually by combining performance analysis and scientific mapping. These results contribute to the central theme of the study on gut microbiota, obesity and diabetes, associated with multivariate analysis from biostatistics and health, red cluster.

Based on the above (Figure 5), the thematic map shows the clusters and *keywords Plus* from 2017 to September 22, 2022 ("recent analysis") identified by the co-occurrence network; The focus of motor issues (1) and emerging topics (3) is Data *Science* (green), where it is possible to demonstrate of great importance the statistical techniques or methods that contribute to the subject of study as multivariate analysis (e.g.: logistic regression and odds ratio). As motor issues, people of both sexes and any age group related to risk factors and retrospective studies continue to be concerned about the increasing occurrence of chronic non-communicable diseases [14-28].

Bibliometrix can analyze keywords, but also terms in article titles and abstracts. Figure 6a shows a conceptual structure map containing a total of 50 keywords that were grouped almost entirely into a single cluster (red) performed by network analysis or multiple correspondence analysis (MCA) in a two-dimensional plot.

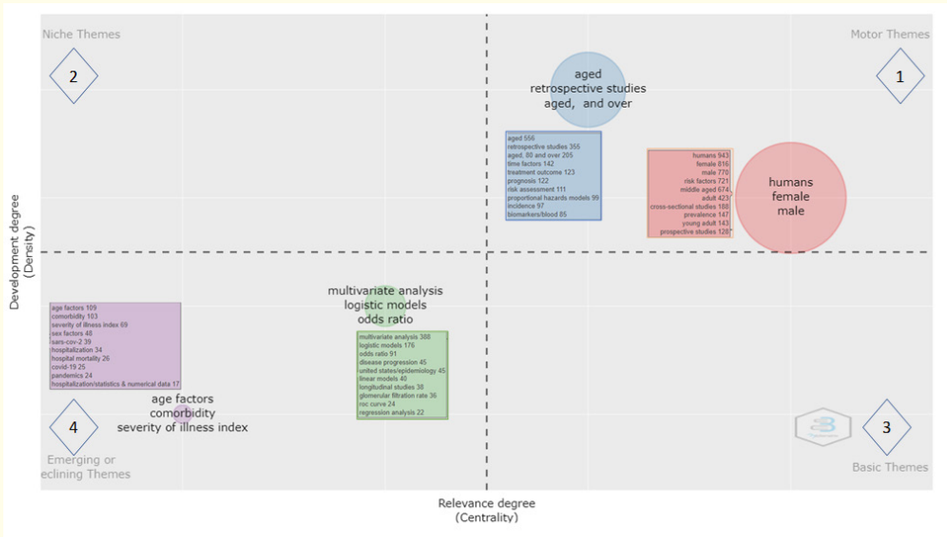


Figure 5: Thematic map. Strategic diagram and results of this work.
Source: Authors' own elaboration based on [8].

68% of the variability of the information is explained in the factorial foreground of the CSF, which is quite high for the average (it explains up to 20% of many of the studies reported for this type of methodology), which indicates a high reliability of the results.

Figure 6b and Table 1 show the top ten (10) authors who contribute the most to plan 1-2 of the CSF (73% of the total): Avgerinos ed, 2017, J vasc surg [1.39%] [29].; Bodewes TCF, 2017, j vasc surg [1.36%] [30].; Ricco JB, 2017, J vasc surg [1.34%] [31].; Stavrou-

Authors	Title	Journal	DOI	Citations
Avgerinos, E; Farber, A; Ali, A; Rybin, D; Doros, G, Eslami, M. [29].	Early carotid endarterectomy performed 2 to 5 days after the onset of neurologic symptoms leads to comparable results to carotid endarterectomy performed at later time points	Journal of Vascular Surgery	10.1016/J.JVS.2017.05.101 [DOI]	24
Bodewes, T; Pothof, A; Darling, J; Deery, S; Jones, D; Soden, P; Moll, F; Schermerhorn, M. [30].	Preoperative anemia associated with adverse outcomes after infrainguinal bypass surgery in patients with chronic limb-threatening ischemia	Journal of Vascular Surgery	10.1016/J.JVS.2017.05.103 [DOI]	22
Rich, J; Gargiulo, M; Star, A; Abualhin, M; Gallitto, E; Desvergnes, M; Schneider, B. [31]	Impact of angiosome- and nonangiosome-targeted peroneal bypass on limb salvage and healing in patients with chronic limb-threatening ischemia	Journal of Vascular Surgery	https://doi.org/10.1016/j.jvs.2017.04.074	29
Stavroulakis, K; Borowski, M. Torsello, G; Bisdas, T. [32].	Association between statin therapy and amputation-free survival in patients with critical limb ischemia in the CRITISCH registry	Journal of Vascular Surgery	https://doi.org/10.1016/j.jvs.2017.05.115.	49
Hong, X; Yes, Q; He, J; Wang, Z; Yang, H; Qi, S; Chen, X; Wang, C; Zhou, H; Li, C; Qin, Z; Xu, F. [33].	Prevalence and clustering of cardiovascular risk factors: a crosssectional survey among Nanjing adults in China	BMJ Open	doi:10.1136/ bmjopen-2017-020530.	25
Morris, N; Estuardo, S; Riley, M; Maguire, G. [34].	Differential Impact of Malnutrition on Health Outcomes Among Indigenous and Non-Indigenous Adults Admitted to Hospital in Regional Australia - A Prospective Cohort Study	Nutrients	https://doi.org/10.3390/nu10050644.	8
Baber U, Chandrasekhar J, Sartori S, et al. [35].	Associations Between Chronic Kidney Disease and Outcomes with Use of Prasugrel Versus Clopidogrel in Patients with Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention	JACC Journal	https://doi.org/10.1016/j.jcin.2017.02.047	43
Vogel, T; Smith, J; Kruse, R. [36].	The association of postoperative glycemic control and lower extremity procedure outcomes.	Journal of Vascular Surgery	https://doi.org/10.1016/j.jvs.2017.01.053	27
Gupta, A; Aridi, H; Locham, S; Nejim, B; Veith, F; Malas, M. [37].	Real-world evidence of superiority of endovascular repair in treating ruptured abdominal aortic aneurysm.	Journal of Vascular Surgery	https://doi.org/10.1016/j.jvs.2017.11.065	30
Reed, G; Young, L; Bagh, I; Maier, M; Shishehbor, M. [38].	Hemodynamic Assessment before and after endovascular therapy for critical limb ischemia and association with clinical outcomes.	JACC Journal	DOI: 10.1016/j.jcin.2017.06.063	22

Table 1: Top 10 - Most Cited References (MCA). Own elaboration.

Figure 6: (a) Conceptual Structure Map; (b) Factor map of the documents with the highest contribution; (c) Thematic Dendogram.

lakis K, 2017, J vasc surg [1.27%] [32].; Hong X, 2018, BMJ Open [1.2%] [33].; Morris N, 2018, Nutrients [1.19%] [34].; Baber U, 2017, JACC cardiovasc interv [1.11%] [35]; Vogel TR, 2017, J Vasc Surg [1.08%] [36].; Gupta AK, 2018, J Vasc Surg [1.05%] [37].; Reed GW, 2017, JACC cardiovasc interv [1.05%] [38].

The Cluster identified groups of documents that express common concepts, the clusters were formed by the level of relationship that coexisted between the related keywords in the metadata of the analyzed documents (Figure 6c).

It is difficult to conduct high-quality data analysis if the expert is not prepared for the discipline to be more data-driven. It is important to have interdisciplinary cooperation and to have an appropriate toolbox to handle data. To a statistician, the terms "machine learning," "predictive modeling," and "big data" refer to statistical methods for analyzing high-dimensional data that are covered in any modern textbook, for example, [39]. The logistic regression model remains an indispensable machine learning tool even for big data analysis and creating multidimensional data visualizations as a means of communicating insights. The academic community has begun to support its methodological developments by directly providing collaborative open-source software tools that implement the methods, e.g., time series modeling and visualization, even having methods for synthesis of epidemiological and genetic data [40], an interdisciplinary approach is needed in light of the increased demand for analytical skills, Cooperation with data and data modeling experts is needed.

- Our contribution from research experience for a multivariate data analysis protocol (e.g., contaminated water in a population of children aged 0-5 years) [1,4,9,10,41].
- From data obtained in electronic records, for description of continuous variables (e.g., viral load, age, weight, height, BMI) mean ± standard deviation, 95% CI for mean and variance.
- To estimate proportions and 95% CI for the presence of pathogens in drinking water or in stool samples from children aged 0-5 years: enteric viruses (Norovirus, Rotavirus and Adenovirus) and protozoa: *Giardia duondenalis*, *Cryptosporidium parvum* and *Entamoeba histolytica.*
- In order to achieve the detection of risk factors in the drinking water of the home (place) or in children aged 0-5 years (individuals), the detection of variables associated (socioeconomic, physiological, clinical history, water management, consumption habits) to the presence or absence of pathogens is proposed, which is achieved through Pearson's Chi-square tests(Proportions, difference of proportions, independence test and/or homogeneity test) and statistical significance is considered if $p < 0.05$ in the bivariate analysis.
- Calculate crude and adjusted ORs to look for an association between the risk of pathogens (enteric viruses and protozoa) and other variables of interest.
- The significant study variables (socio-economic, physiological, clinical history, water management, consumption habits) will be incorporated into a multivariate logistic regression model, taking as a dependent variable the presence (1) or absence (0) in each study pathogen for the water samples for human consumption or in each study protozoan of the fecal samples of children.
- To fit a log-linear model with significant effects of the variables of interest on the risk of developing pathogens in children under 5 years of age, which showed correct classification rate, sensitivity and specificity, and a capacity for discrimination assessed through the area of the OCR curve.
- To perform Multiple Correspondence Analysis to relate socioeconomic and physiological variables, clinical history, water management, and consumption habits of the child population in the department of Sucre with the prevalence of pathogens in water for human consumption or protozoa in fecal samples of evaluated children.
- Using results of pathogens in drinking water and protozoa in fecal samples from the children under study, this information is correlated by means of multiple factorial multivariate descriptive analysis, where the unit of study will be the home.
- Perform simple correspondence analyses (SCAs) for the reading of contingency tables by cross-referencing qualitative study variables with pathogens, in order to characterize risk factors in children under 5 years of age.
- Relate anthropometric variables (Weight, Height, Age, BMI) with the different protozoa present in feces (Canonical Correspondence Analysis).
- Perform a Monte Carlo simulation to obtain distributions of probability of infection and burden of disease, estimating the probability of infection by using the available information on what is known as "minimum infectious dose" (< ID50). The burden of disease will be obtained by multiplying the annual probability of infection and the burden of disease caused by an infection.
- The statistical analyses described above will be performed with the R statistical software.

Our bibliometric work revealed that there is not yet enough evidence on which to base recommendations for the selection of variables and functional forms in multivariate analysis. Such evidence may come from comparisons between alternative methods. In particular, we highlight seven important topics that require further research for the direction of future research: Investigation and comparison of the properties of variable selection strategies; Comparison of spline procedures in a univariate and multivariate context; How to model one or more variables with a 'zero peak'? Comparison of multivariate procedures for the selection of models and functions; Role of shrinkage in correcting bias introduced by data-dependent modeling; Evaluation of new approaches to post-selection inference; Is it necessary to adapt procedures for very large sample sizes?

Conclusions

This bibliometric study suggests that statistical analyses that are performed with multivariate analysis (logistic regression, multiple correspondence analysis, Kaplan and Meier method, PCA, epidemiology), studies show that you have a lot to gain by bringing a statistician on board data-driven projects from the start.

Among the risk factors are age and gender, COVID19 had a high mortality rate in people with obesity and diabetes, among the novelties dietary intake has shifted towards a high-fat, high-car-

bohydrate and low-fiber diet that may have resulted in functional changes in gut microbiota.

Among the microbiological organisms, *Mycobacterium tuberculosis* (TB), *Clostridium difficile* (CDI) and *Candida vaginalla* are shown as possible risks for the presence of diseases such as overweight, obesity and diabetes.

Thanks

Thanks to the open science and open source community in carrying out this research.

Bibliography

1. Vertel-Morinson M., *et al*. "Rural Education in Sucre - Dissertation from Multivariate Analysis". Editorial Universidad de Sucre (2018).

2. Kawuki J., *et al*. "A bibliometric analysis of childhood obesity research from China indexed in Web of Science". *Journal of Public Health and Emergency* 5 (2021): 3.

3. Rousseau DM. "The Oxford handbook of evidence-based management". Oxford University Press (2012).

4. Vertel-Morinson M., *et al*. "Sociodemographic and Parasitological Factors Determining Learning Capacity and Nutritional Status in Rural Schoolchildren: Data Mining for Decision Making". *Acta Scientific Nutritional Health* 6.12 (2022): 15-22.

5. Trueba-Gómez R., *et al*. "The PubMed database and the search for scientific information". *Seminarios de la Fundación Española de Reumatología* 11 (2010): 49-63.

6. Aria M and Cuccurullo C. "*Bibliometrix*: An R-tool for comprehensive science mapping análisis". *Journal of Informetrics* 11 (2011): 959-975.

7. Newman ME. "Scientific collaboration networks. I. Network construction and fundamental results". *Physical Review E* 64.1 (2001): 016131.

8. Cobo MJ., *et al*. "SciMAT: A new science mapping analysis software tool". *Journal of the American Society for Information Science and Technology* 63.8 (2012): 1609-1630.

9. Vertel M., *et al*. "Multivariate Data Analysis. Application: Dual Purpose Production System". Ediciones Universidad Simón Bolívar, a publishing house recognized by COLCIENCIAS. 223 Book: research result (2016).

10. Vertel M and Pardo C-E. "Comparison between canonical correspondence analysis and multiple factor analysis in continuous frequencies-variable tables". *Master's thesis*, Universidad Nacional de Colombia, Bogotá (2010).

11. Dray S., *et al*. "Co-inertia analysis and the linking of ecological data tables". *Ecology* 84.11 (2003): 3078-3089.

12. Moral-Muñoz JA., *et al*. "Software tools for conducting bibliometric analysis in science: an up-to-date review". *El Profesional de la Información* 29.1 (2020): e290103.

13. R Core Team. "R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing (2021).

14. Chen C. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literatura". *Journal of the Association for Information Science and Technology* 57 (2006): 359-377.

15. Chung Y., *et al*. "Serum cystatin C is associated with subclinical aterosclerosis in patients with type 2 diabetes: A restrospective study" (2017).

16. Chandradas S Khalili., *et al*. "Does Obesity Influence the Risk of Clostridium difficile Infection Among Patients with Ulcerative Colitis?" *Digestive Diseases and Sciences* 63.9 (2018): 2445-2450.

17. Llanera D Wilmington., *et al*. "Clinical Characteristics of COVID-19 Patients in a Regional Population With Diabetes Mellitus: The ACCREDIT Study. *Frontiers in Endocrinology* 13 January 2022. Sec. Clinical Diabetes. This article is part of the Research Topic. Covid-19 and Diabetes - Volume II (2022).

18. Ávila EE., *et al*. "High Relative Abundance of Lactobacillus reuteri and Fructose Intake are Associated with Adiposity and Cardiometabolic Risk Factors in Children from Mexico City". *Nutrients* 11 (2019): 1207.

19. Stanislawski M., *et al*. "Gut microbiota in the first 2 years of life and the association with body mass index at age 12 in a Norwegian birth cohort". *mBio* 9 (2018): e01751-18.

20. Mirpuri J. "Evidence for maternal diet-mediated effects on the offspring microbiome and immunity: implications for public health initiatives". *Pediatric Research* 89.2 (2020): 301-306.

21. Herrera A and López M. "Childhood obesity: current situation in Mexic". *Frontiers in Public Health* 10 (2018): 949893.

22. Needell JIr., *et al*. "Maternal treatment with short-chain fatty acids modulates the intestinal microbiota and immunity and ameliorates type 1 diabetes in the offspring". *PLOS One* (2017).

23. Fu CP Lee., *et al*. "Metformin as a potential protective therapy against tuberculosis in patients with diabetes mellitus: A retrospective cohort study in a single teaching hospital". *Journal of Diabetes Investigation* 12.9 (2021): 1603-1609.

24. Ndahayo Sophonie., *et al*. "Risk-Factors Associated with Overweight and Obesity Among Adolescents in Selected Urban and Peri-Urban Secondary Schools in Monze, Zambia". *Acta Scientific Nutritional Health* 6.8 (2022).

25. VanWagner L., *et al*. "Alcohol use and cardiovascular disease risk in patients with nonalcoholic fatty liver disease". *Gastroenterology* 153.5 (2017): 1260-1272.

26. Van Eck NJ and Waltman L. "Software survey: VOSviewer, a computer program for bibliometric mapping". *Scientometrics* 84 (2010): 523-538.

27. Wang Z., *et al*. "Microbial co-occurrence complicates associations of gut microbiome with US immigration, dietary intake and obesity". *Genome Biology* 22 (2021): 336.

28. Yokoyama H Naga., *et al*. "Incidence and risk of vaginal candidiasis associated with sodium-glucose cotransporter 2 inhibitors in real-world practice for women with type 2 diabetes". *Journal of Diabetes Investigation* 10.2 (2018): 439-445.

29. Avgerinos., *et al*. "Early carotid endarterectomy performed 2 to 5 days after the onset of neurologic symptoms leads to comparable results to carotid endarterectomy performed at later time points". *Journal of Vascular Surgery* 66.5 (2017): 1719-1726.

30. Bodewes., *et al*. "Preoperative anemia associated with adverse outcomes after infrainguinal bypass surgery in patients with chronic limb-threatening ischemia". *Journal of Vascular Surgery* 66.6 (2017): 1775-1785.

31. Ricco J., *et al*. "Impact of angiosome- and nonangiosome-targeted peroneal bypass on limb salvage and healing in patients with chronic limb-threatening ischemia". *Journal of Vascular Surgery* 66.5 (2017): 1479-1487.

32. Stavroulakis K., *et al*. "Association between statin therapy and amputation-free survival in patients with critical limb ischemia in the CRITISCH registry". *Journal of Vascular Surgery* 66.5 (2017): 1534-1542.

33. Hong X., *et al*. "Prevalence and clustering of cardiovascular risk factors: a crosssectional survey among Nanjing adults in China (2018).

34. Morris N., *et al*. "Differential Impact of Malnutrition on Health Outcomes Among Indigenous and Non-Indigenous Adults Admitted to Hospital in Regional Australia-A Prospective Cohort Study (2018).

35. Baber U., *et al*. "The association of postoperative glycemic control and lower extremity procedure outcomes (2018).

36. Gupta A., *et al*. "Real-world evidence of superiority of endovascular repair in treating ruptured abdominal aortic anerysm (2018).

37. Reed G., *et al*. "Hemodynamic Assessment before and after endovascular therapy for critical limb ischemia and association with clinical outcomes (2017).

38. Efron B and Hastie T. "Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press (2021).

39. Jombart T., *et al*. "Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data". *PLoS Computational Biology* 10.1 (2014): e1003457.

40. Vertel M., *et al*. "Multivariate analysis of the quality of education in Sucre". *Scientia et Technica* 19.1 (2014): 96-105.