Bibliometric Analysis of Research on Firth Penalized Logistic Regression in Addressing Complete Separation

Nurul Husna Jamian[1*], Ahmad Zia Ul-Saufie[2], Mohammad Nasir Abdullah[3]

[1,3] *Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia.*

[2] *Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia.*

ARTICLE INFO

Article history:
Received 26 May 2025
Revised 18 August 2025
Accepted 19 August 2025
Online first
Published 1 September 2025

Keywords:
Firth Penalized Logistic Regression
Complete Separation
Logistic Regression
Bibliometric Analysis

DOI:
10.24191/jcrinn.v9i2.535

ABSTRACT

Complete separation in logistic regression leads to infinite estimates which prevents reliable inference. Firth's penalized likelihood method has emerged as a widely accepted and reliable solution that provides finite and more stable estimates. Despite its growing relevance, a thorough understanding of the global research on this topic remains limited. This study conducts a bibliometric analysis of trends related to complete separation in logistic regression using Firth penalized regression. Bibliographic data were retrieved from the Scopus database and analysed using Microsoft Excel and VOSviewer software. After applying inclusion criteria, nine journal articles published between 2012 and 2024 were identified through a structured search conducted on February 22, 2025. The findings reveal a small but growing body of literature, reflecting the emerging status of research on complete separation in logistic regression using Firth penalized regression. The results show an upward trend in publications, particularly from 2019 onward with the United States and Malaysia identified as the most productive countries. Influential articles contributed to methodological development and applications in health and transportation research. Keyword co-occurrence analysis identified thematic clusters in human studies, statistical modelling, and estimation techniques. These findings provide an overview of publication trends, collaboration networks, and research gaps which could support future methodological and multidisciplinary integration of Firth penalized regression.

1. INTRODUCTION

Logistic regression is a widely used statistical method for modelling binary outcome variables, particularly in fields such as medicine, epidemiology, social sciences, and genetics (Gilbert, 2022; Hess & Hess, 2019). It is commonly applied to predict outcomes such as disease occurrence, patient readmission, or treatment efficacy (Gilbert, 2022). The technique has shown impressive results in medical research, aiding in the identification of risk factors, assessment of disease probabilities, and support for clinical decision-making (Olowe et al., 2024). Logistic regression models apply concepts such as odds ratios and logit transformation to examine relationships between predictors and outcomes (Nathanson & Higgins, 2008).

However, logistic regression models can encounter serious estimation issues when the data exhibit a phenomenon known as complete separation. Complete separation arises when a predictor or combination of predictors perfectly discriminates between outcome categories, causing the maximum likelihood estimates (MLE) of regression coefficients to diverge towards infinity (Allison, 2008). This problem compromises the reliability of model interpretation, standard errors, and inference, rendering the traditional logistic regression model unsuitable for such datasets (Botes & Fletcher, 2014).

[1*] Corresponding author. *E-mail address*: nurul872@uitm.edu.my
https://doi.org/10.24191/jcrinn.v10i2.535

To address this challenge, Firth (1993) proposed a penalized likelihood approach that applies a bias-reducing modification to the score function. This method known as Firth's penalized logistic regression that has been demonstrated to effectively resolve issues arising from complete separation without the need to remove covariates (Clark et al., 2023). Firth's approach quickly gained traction due to its ability to produce finite, stable, and more reliable parameter estimates, especially in small sample sizes or sparse data conditions. These characteristics make it particularly useful in fields such as genetics, rare disease research, and case-control study designs (Suhas et al., 2023; D'Angelo & Ran, 2024).

Given its rising prominence and methodological importance, a bibliometric analysis is both timely and warranted. Bibliometric analysis is a quantitative method used to assess academic literature within a specific field, providing a comprehensive overview of publication trends, influential works, author and institutional productivity, collaboration networks, and thematic developments over time (Kumar, 2025). Prior studies have highlighted its methodological strengths, little is known about its bibliometric landscape that is, where, how, and by whom the method has been applied and studied. By examining publication patterns, citation structures, and keyword trends, researchers can identify research gaps, emerging themes, and identify potential areas for future exploration.

The objectives of this study are to conduct a bibliometric analysis of global research related to complete separation in logistic regression using Firth penalized regression. Specifically, the study aims to examine publication trends over time to understand the growth and evolution of interest in this topic. In addition, it intends to identify the most influential articles, authors, and journals. Through this approach, the study seeks to highlight research patterns, shifts, and contributions within the literature, providing a comprehensive understanding of the development and impact of Firth penalized regression in addressing complete separation in logistic regression models.

2. LITERATURE REVIEW

Complete separation in logistic regression occurs when one or more independent variables perfectly predict the binary outcome, leading to infinite or zero maximum likelihood estimates (Botes & Fletcher, 2014; Mansournia et al., 2017). This issue presents a significant challenge, particularly prevalent in small samples or datasets with rare events (Mansournia et al., 2017). Traditional maximum likelihood estimation (MLE) often fails under these conditions, leading to infinite parameter estimates. Firth's penalized likelihood method has been widely adopted to address this issue, offering finite and more reliable estimates (Alam et al., 2022).

Firth (1993) laid the foundational work by introducing a general bias-reduction method for MLEs through penalized likelihood. Although originally proposed for exponential family models, its application in logistic regression proved particularly impactful. The technique modifies the score function to counteract the first-order bias of MLE, thereby producing finite and less biased estimates. Subsequent research by Heinze and Schemper (2002) adapted Firth's approach specifically to logistic regression, demonstrating its superiority over traditional methods when complete or quasi-complete separation occurs. Their simulations revealed that Firth's method not only yields finite estimates but also improved coverage probabilities for confidence intervals, particularly in small-sample settings.

Puhr et al. (2017) evaluated Firth's logistic regression under rare event conditions and found that although it effectively reduced bias in parameter estimates, it tended to skew predicted probabilities toward 0.5 in imbalanced datasets. To address it, they proposed post-hoc and augmentation-based corrections to improve predictive accuracy without undermining bias reduction. Walker and Smith (2019) found that sparseness in binary predictors introduces substantial bias with small sample sizes, which Firth's procedure can effectively correct. Karabon (2020) emphasized Firth's method as a solution for rare events and complete separation, demonstrating its advantages over other approaches such as Fisher's Exact test and Exact logistic regression. Stolte et al. (2024) provided a broader review of bias-reduction methods, stating

https://doi.org/10.24191/jcrinn.v10i2.535

that Firth's estimator often performs best across circumstances. Rainey and McCaskey (2021) demonstrated that the penalized maximum likelihood estimator reduces both bias and variance when compared to standard MLE, yielding significant improvements in small samples (e.g., 50 observations) and perceptible gains even in larger datasets (e.g., 1,000 observations). Zorn (2005) supported the use of Firth's logistic regression in political science, stressing its advantages for stratified datasets prone to separation. While recognizing its effectiveness, he observed that its application remained limited beyond biostatistics, indicating a need for broader disciplinary dissemination.

Collectively, the scholarly contributions establish a strong foundation for the current study. They illustrate the development and application of Firth penalized regression, its empirical advantages, and its disciplinary adoption. However, several gaps persist in the literature (De Oliveira et al., 2019). First, while the methodological effectiveness of Firth's correction is well-documented, the bibliographic and thematic evolution of its usage remains uncharted. Second, little is known about the global research network, citation patterns, and institutional contributions to this domain. Lastly, despite the increasing volume of studies employing Firth regression, no bibliometric study has yet attempted to comprehensively map the literature. The present study addresses these gaps by conducting a detailed bibliometric analysis of scholarly works on complete separation in logistic regression using Firth's penalized method. This study uses bibliometric tools and approaches to find publication trends, influential authors, leading journals, collaboration patterns, and emerging themes. This provides a meta-perspective on how the area has evolved and where future research prospects exist. The literature covered here not only informs the conceptual and methodological foundations of this work, but it also demonstrates its importance in furthering understanding of the scholarly effect and trend of Firth penalized regression. Fig. 1 depicts the phases of bibliometric analysis.

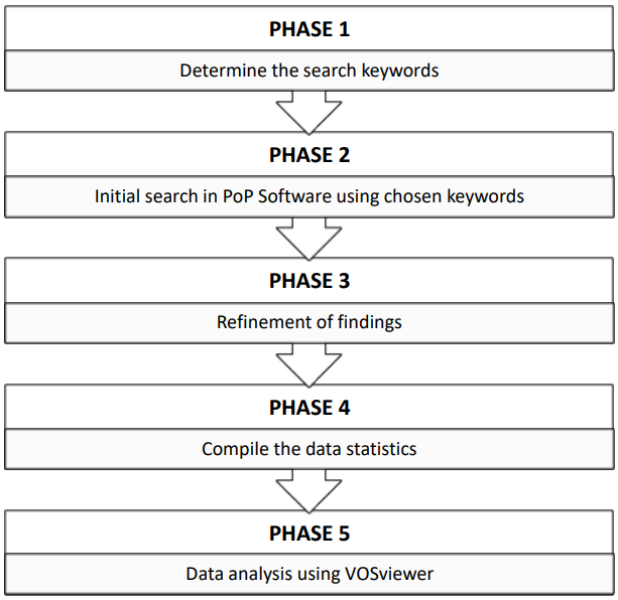

Fig. 1. The phase of bibliometric analysis of shift scheduling

Source: Sarudin et al. (2023)

3. METHODOLOGY

This study retrieved bibliographic data from the Scopus database on February 22, 2025. Scopus was selected as the primary database because it is one of the largest and most comprehensive abstract and

citation databases of peer-reviewed literature. It is widely recognized for its extensive multidisciplinary coverage and is frequently used in bibliometric and literature review studies (Rojas-Flores et al., 2023; Sarudin et al., 2024). Compared to other digital databases, Scopus has been found to offer the most diverse field coverage (Yaman et al., 2019). To ensure data integrity, records were cross-checked for duplicate entries by comparing titles, authors, and publication years.

Bibliographic data can also be retrieved using software such as Publish or Perish, developed by Anne-Will Harzing (Harzing, 2023). However, it imposes a limit of 1,000 records per query (Sarudin et al., 2023). This constraint reduces its suitability for comprehensive bibliometric reviews. Scopus, in contrast, allows for larger and more flexible data retrieval (Baas et al., 2020). Although Dimensions has emerged as a potential alternative to Scopus and Web of Science and often provides similar coverage (Harzing, 2019; Martín-Martín et al., 2021), it still showed reduced yield for this specific topic. Google Scholar was also excluded due to reproducibility limitations. In contrast, Scopus offers more stable and reproducible results across regions (Pozsgai et al., 2020), which is critical for ensuring the transparency and reliability of bibliometric analyses.

The data are obtained using the topic search query, as illustrated in Fig. 2. The initial search string used was TITLE-ABS-KEY ((“complete separation" OR "separation issue”) AND (“logistic regression”) AND (“Firth" OR "Firth's correction" OR "penalized regression”)). The results then underwent a filtering process based on several criteria including document type, year of publication, and language. The refined Scopus search string was including document type, year of publication, and language. The refined Scopus search string was TITLE-ABS-KEY (("complete separation" OR "separation issue") AND ("logistic regression") AND ("Firth" OR "Firth's correction" OR "penalized regression")) AND (LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2023)) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")).

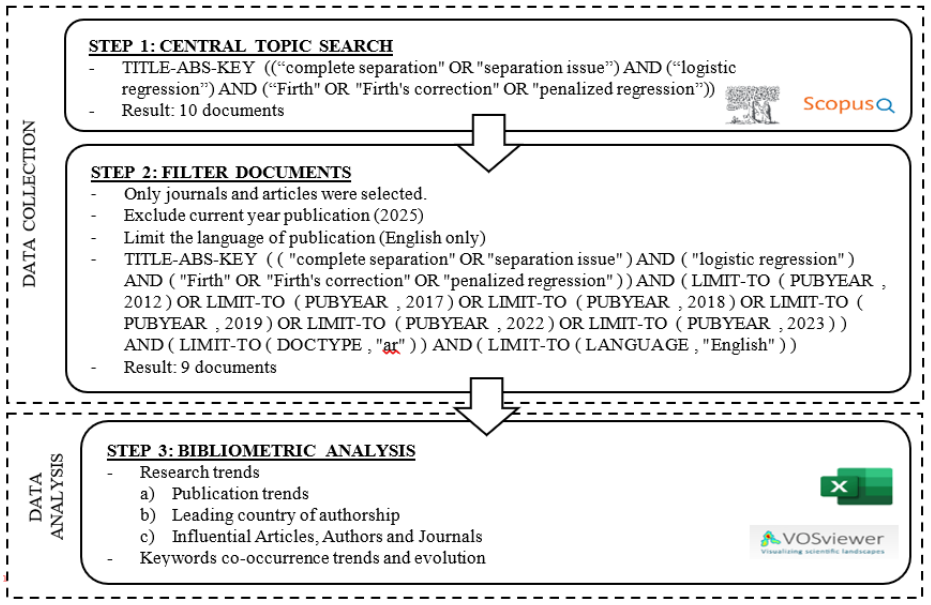

Fig. 2. Workflow of data retrieval and analysis for the bibliometric review

The inclusion criteria consisted of journal articles journal articles published in English between 2012 and 2024. Only English-language publications were included in the analysis to ensure consistency in text-mining, keyword analysis, and citation matching, as non-English articles often lack standardized metadata

in bibliographic databases. In addition, multiple investigations suggest that restricting reviews to English-language publications has minimal effect on overall conclusions (Nussbaumer-Streit et al., 2020; Dobrescu et al., 2021).

The remaining documents are then analyzed to determine research trends in leading countries. The analysis also identifies prominent authors and journals that actively publish articles related to complete separation in logistic regression using Firth penalized regression. Additionally, frequently used keywords from previous research were examined. Initially, ten documents on this topic were identified in the Scopus database, covering the period from 2012 to 2025. Of these, nine articles met the inclusion criteria of being journal articles published in English up to the year 2024 as suggested by Sarudin et al. (2023) and Ilmasari et al. (2022). These articles were then exported in Comma-Separated Values (CSV) format for bibliometric analysis.

This study employs several analytical tools to analyze and visualize the findings. VOSviewer will be used to conduct country co-authorship analysis and keyword co-occurrence analyses. It will also illustrate the results through network visualizations that show the connections between documents, authors, or countries in the dataset (Van Eck & Waltman, 2023; Ilmasari et al., 2022). The stronger the correlation between two nodes, the higher the link strength assigned (Ilmasari et al., 2022). According to Perianes-Rodriguez et al. (2016), VOSviewer offers two counting methods: full counting and fractional counting.

In VOSviewer, the country co-authorship analysis was conducted using full counting, where each co-author's country receives full credit for a publication regardless of the number of co-authors from other countries. Full counting is straightforward to interpret and provides an intuitive measure of total participation, making it particularly suitable for identifying absolute productivity and visibility of countries in a small dataset (Iskandar et al., 2020). An alternative is fractional counting, where each publication's credit is divided proportionally among co-authors' countries. It offers a more precise representation of relative contribution. It can underrepresent the role of countries engaged in large and highly collaborative projects (Donner, 2020).

In this study, the dataset contains nine publications and collaboration patterns are relatively sparse, full counting was preferred as it avoids diminishing the visibility of countries involved in multi-country papers. Besides, it assigns integer values to link strength based on the number of co-authored documents (Van Eck & Waltman, 2023). In addition to VOSviewer, Microsoft Excel will be used to compute data metrics and generate table.

4. FINDINGS

This bibliometric analysis presents results in two subsections: (1) research trends and (2) keyword evolution across chronological groupings. The findings aim to help future researchers understand the development of this area and identify opportunities for new topics or algorithms.

4.1 Research trends

The research trends related to complete separation in logistic regression using Firth penalized regression are examined in this section. The discussion covers publication trends, authorship by country, and influential articles, authors, and journals.

4.1.1 Publication Trends

Based on data obtained from the Scopus database, the number of publications on the topic of complete separation in logistic regression using Firth penalized regression has shown significant growth particularly over the past decade. This trend reflects a growing interest and ongoing advancements in the field. A total of ten documents published between 2012 and 2025 were identified where the topic appeared in the title,

https://doi.org/10.24191/jcrinn.v10i2.535

abstract, or keywords. Among these, journal articles were the most common publication type, accounting for all ten documents. After the screening phase, only English-language journal articles published up to the year 2024 were selected, resulting in a final dataset of nine journal articles. All nine articles included in this study are published in peer-reviewed, properly archived journals indexed in Scopus, ensuring scholarly quality and long-term accessibility. This improves the reliability of the sources and the strength of the bibliometric mapping.

As shown in Fig. 3, the highest number of publications occurred in 2023, 2022, and 2019, with two articles each. In contrast, only one article was published in 2018, 2017, and 2012. An analysis of the publication trends on complete separation in logistic regression using Firth penalized regression reveals a clear pattern across two distinct phases. In the initial phase (2012–2018), only three articles were published, accounting for 33.33% of the total. In the development phase (2019–2024), the number increased to six articles, representing 66.67% of the publications. This growth indicates an increasing research interest and a broader exploration of methodologies, indicating significant progress in the field.

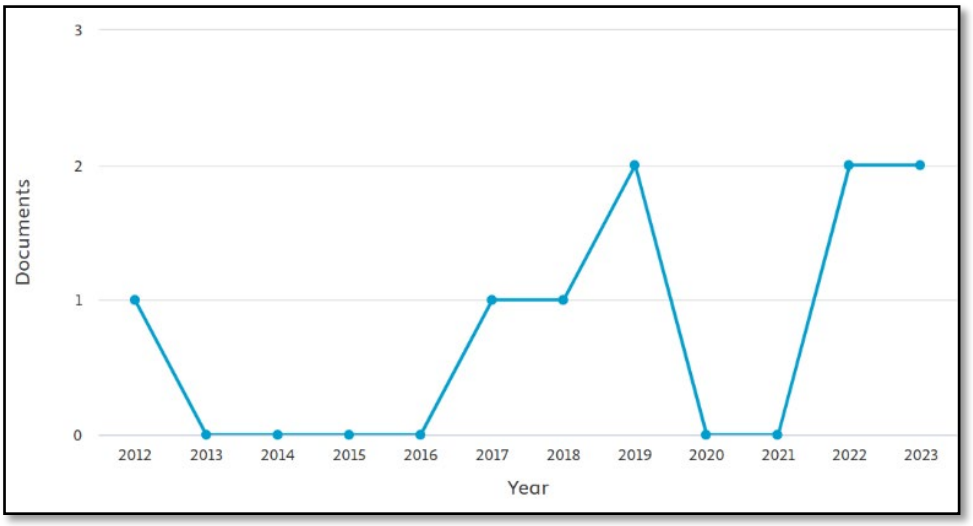

Fig. 3. Annual publication trends based on Scopus data, 2012–2024

4.1.2 Leading country of authorship

The geographical distribution of research on complete separation in logistic regression using Firth penalized regression demonstrates worldwide interest in this study area and highlights collaborative efforts. This subsection examines contributions from various countries, focusing on the leading countries with the most articles published in the Scopus database. VOSviewer was used to analyze and visualize international collaboration through network maps. The co-authorship analysis was conducted with full counting, considering only countries with at least one article and a maximum of 25 countries per document.

As a result, eight countries published articles on complete separation in logistic regression using Firth penalized regression between 2012 and 2024. The United States contributed two articles (Noor & Asmael, 2023; Choi et al., 2018), while Malaysia also produced two articles (Abdullah et al., 2022a; Abdullah et al., 2022b), making them the most productive countries in this area. Malaysia began its contributions in 2022, while the United States started in 2018. In addition, Slovenia, India, South Korea, Poland, Iraq, and Australia each contributed one article. Beyond focusing the leading countries, this study also explores

international collaboration in this research domain. International collaborations among these eight countries were analyzed using VOSviewer and are visualized in a network map (Fig. 4). All eight countries met the inclusion threshold at least one document per country and a maximum of 25 countries per document.

In the network, the country co-authorship network based on the identified publications. Each node represents a country, and the spatial distribution suggests limited collaboration among countries as no visible links (edges) are present between the nodes. This indicates that the publications in the dataset were largely authored within single countries without international co-authorship. Countries such as the United States and Malaysia appear more prominently, which may reflect a higher number of publications rather than actual collaboration. In contrast, countries like Iraq, India, Poland, and South Korea are represented as isolated nodes, suggesting opportunities for strengthening international research ties in this area.

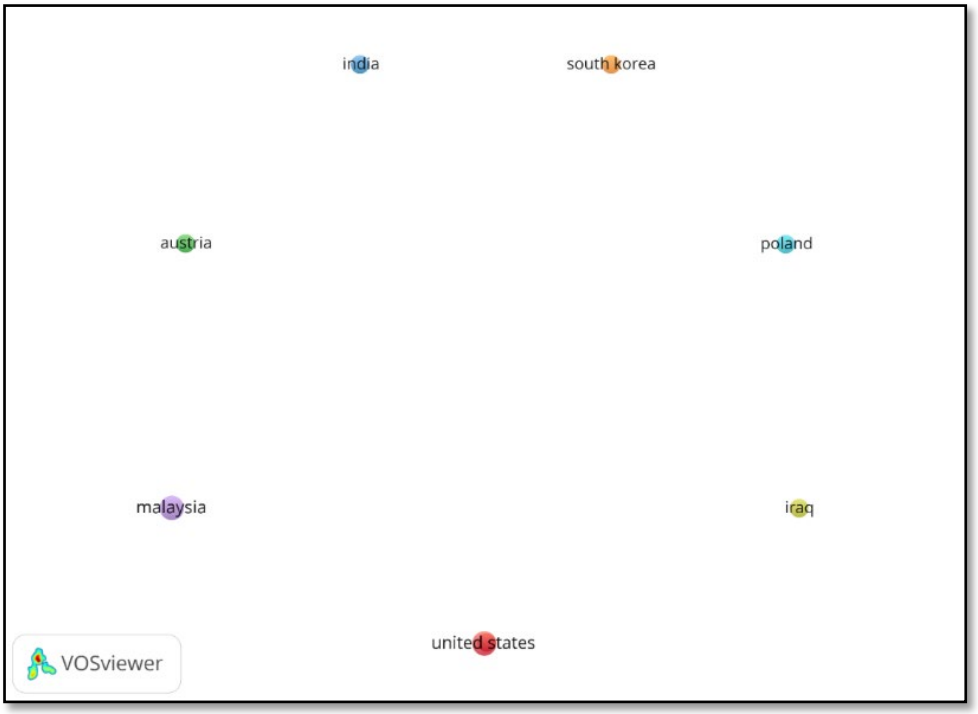

Fig. 4. Network visualization map of international collaboration on complete separation in logistic regression using Firth penalized regression research from 2012 to 2024.

Source: Online map: https://tinyurl.com/27vfxvfs

These countries were clustered into groups based on the frequency of collaboration using the Louvain algorithm for modularity optimization (Van Eck & Waltman, 2017; Blondel et al., 2008). This method detects communities by maximizing modularity, grouping together countries that collaborate more frequently with each other than with those outside the group. The number of clusters is not pre-defined. It is determined automatically by the algorithm based on the underlying network structure (Blondel et al., 2024; Rahiminejad et al., 2019).

In this dataset, the algorithm produced seven distinct clusters which reflects the relatively sparse collaboration patterns among countries in this research area. Countries that collaborate frequently tend to have similar co-authorship patterns (Sarudin et al., 2024). The clustering results were as follows: Cluster 1

(red) included Australia and the United States, indicating academic collaboration between them. Cluster 2 (green) consisted of Austria, Cluster 3 (blue) included only India, Cluster 4 (yellow) included Iraq, Cluster 5 (purple) consisted of Malaysia, Cluster 6 (sea blue) included Poland, and Cluster 7 (orange) consisted of South Korea.

4.1.3 Influential articles, authors, and journals

Table 1 presents the ranking of articles related to complete separation in logistic regression using Firth penalized regression based on the total number of citations. The highest-ranked article is "Maximum Likelihood and Firth Logistic Regression of the Pedestrian Route Choice," published in 2017 in the International Regional Science Review with a total of 39 citations, contributing approximately 38.61% to the overall citation count. This study authored by Gim and Ko (2016), suggests a significant impact in this research area. In the context of specialized statistical methodology studies where speciality topics often attract fewer total citations than broader applied research, accumulating nearly 40 citations over seven years can be considered relatively high.

The second highest-ranked article is "Bring more data! — A good advice? Removing separation in logistic regression by increasing sample size," published in 2019 in the International Journal of Environmental Research and Public Health with 18 citations and approximately 17.82% contribution. This article authored by Šinkovec et al. (2019), also demonstrates substantial influence on subsequent research. Other significant contributions include the article "Firth's penalized logistic regression: A superior approach for analysis of data from India's National Mental Health Survey, 2016," published in 2023 in the Indian Journal of Psychiatry with 13 citations and about 12.87% contribution of the total, authored by Suhas et al. (2023).

Another significant article is "Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects," published in 2018 in Bioinformatics with 12 citations and approximately 11.88% contribution. It was authored by Choi et al. (2018). Interestingly, several recent articles have already gained considerable attention, suggesting the growing relevance of research on complete separation in logistic regression. Conversely, the article "A Study on Interstate Freight Mode Choice Between Trucks and Trains Used to Transport Oil Products: A Case Study of Iraq," published in 2023 in Transport Problems by Noor and Asmael (2023) has yet to accumulate citations at the time of analysis. This may be due to its recent publication date or narrower topical scope. It may require more time to gain recognition within the academic community.

Some journals within the dataset hold particularly strong influence in their respective fields. For example, Bioinformatics is a leading outlet in computational biology and bioinformatics, with a 2024 Impact Factor (IF) of 5.4 (Q1) and an SCImago Journal Rank (SJR) of approximately 2.45. These metrics reflect its high visibility and academic prominence. Similarly, the Journal of Statistical Software is widely regarded in the field of statistical methodology, with a 2024–2025 Impact Score of approximately 5.8, an SJR of 2.72, and a Q1 ranking, underscoring its methodological impact and widespread citation.

Other journals, such as the International Journal of Environmental Research and Public Health, Australian Veterinary Journal, and International Regional Science Review, also demonstrate moderate to strong standing in their respective domains, typically ranking in Q2 or Q3 quartiles. This mix of high- and mid-tier outlets indicates that the selected articles are drawn from both methodologically influential and applied research venues, providing balanced perspective on the scholarly landscape of research on complete separation in logistic regression using Firth penalized regression.

Table 1. The most influential article on complete separation in logistic regression using Firth penalized regression

No.	Title of article	Year	Journal	Total citations	Contribution rate (%)

1.	Maximum Likelihood and Firth Logistic Regression of the Pedestrian Route Choice (Gim & Ko, 2016)	2017	International Regional Science Review	39	38.61
2.	Bring More Data! —A Good Advice? Removing Separation in Logistic Regression by Increasing Sample Size (Šinkovec et al., 2019)	2019	International Journal of Environmental Research and Public Health	18	17.82
3.	Firth's Penalized Logistic Regression: A Superior Approach for Analysis of Data from India's National Mental Health Survey, 2016 (Suhas et al., 2023)	2023	Indian Journal of Psychiatry	13	12.87
4.	Evaluating Statistical Approaches to Leverage Large Clinical Datasets for Uncovering Therapeutic and Adverse Medication Effects (Choi et al., 2018)	2018	Bioinformatics	12	11.88
5.	Separation-Resistant and Bias-Reduced Logistic Regression: STATISTICA Macro (Fijorek & Sokolowski, 2012)	2012	Journal of Statistical Software	10	9.90
6.	Identification of Blood-Based Multi-Omics Biomarkers for Alzheimer's Disease Using Firth's Logistic Regression (Abdullah et al., 2022a)	2022	Pertanika Journal of Science and Technology	6	5.94
7.	A Review of 91 Canine and Feline Red-Bellied Black Snake (Pseudechis Porphyriacus) Envenomation Cases and Lessons for Improved Management (Wun et al., 2022)	2022	Australian Veterinary Journal	2	1.98
8	Discovering Potential Blood-Based Cytokine Biomarkers for Alzheimer's Disease Using Firth Logistic Regression (Abdullah et al., 2022b)	2019	Epidemiology Biostatistics and Public Health	1	0.99
9.	A Study on Interstate Freight Mode Choice Between Trucks and Trains Used to Transport Oil Products: A Case Study of Iraq (Noor & Asmael, 2023)	2023	Transport Problems	0	0.00

4.2 Keyword co-occurrence trends and evolution

The analysis explores the keyword co-occurrences in research on complete separation in logistic regression using VOSviewer. The analysis included all keywords identified by the author or indexed in the database. This clustering facilitates the identification of thematic groupings within the field, outlining the primary research areas and methodological focus present in the existing literature. Only documents containing at least two occurrences of a keyword were considered. As a result, 12 out of 151 keywords met the threshold and were visualized in Fig. 3. These twelve keywords were grouped into three distinct clusters as shown in Table 2.

In VOSviewer, total link strength (TLS) represents the overall strength of co-occurrence links between a given keyword (node) and all other keywords in the network. A higher TLS indicates that a keyword has stronger or more frequent associations with other keywords, reflecting its centrality and importance in the research field. TLS is calculated by summing the co-occurrence frequencies of the keyword with all others in the dataset (Van Eck & Waltman, 2017). In this study, keywords with higher TLS are considered more influential within the network as they are more strongly connected to multiple thematic areas.

The keyword "human" being the most frequently occurring keyword with three occurrences and a TLS of 13. This was followed by "logistic regression analysis" and "article," both with three occurrences and a TLS of 11. Additionally, "logistic models," "statistical model," and "humans" each appeared twice, also with a TLS of 11. This highlights that research on complete separation in logistic regression most often focuses on applications involving human studies.

The prominence of the keyword "human" indicates that research on complete separation in logistic regression is most commonly applied in studies involving human subjects. This is consistent with its

relevance in biomedical, psychological, and clinical research where issues of separation frequently arise due to rare outcomes or sparse data structures (Mansournia et al., 2017). Such conditions make traditional logistic regression unreliable, thereby necessitating bias-reduction methods such as Firth's penalized likelihood. This observation is further supported by findings that Firth's correction is particularly effective in small or imbalanced human-subject datasets (Alam et al., 2022). These findings highlight the potential for broader adoption of this method in health-related applications.

Cluster 1 which exhibit strong co-occurrence links to both methodological and application-related terms, giving it high degree centrality in the network. It contains six keywords: "human," "logistic regression analysis," "article," "logistic models," "statistical model," and "humans". "Human" stands out as the most influential keyword with three occurrences and a high TLS of 13, showing strong connections with other terms in the cluster. Both "logistic regression analysis" and "article" follow closely with strong connections, each with three occurrences and a TLS of 11. The keywords "logistic models," "statistical model," and "humans" each occur twice and also have a TLS of 11, indicating their significant in the research network.

Cluster 2 includes four keywords: "biomarkers," "complete separation," "Firth logistic regression," and "logistic regression". Within this group, "Firth logistic regression" and "complete separation" are closely linked, highlighting their relevance in addressing separation issues in logistic regression models. Cluster 3 contains two keywords: "regression analysis" and "estimation method". This cluster emphasizes the methodological aspects of logistic regression, focusing on estimation techniques and analytical methods.

Table 2. Clusters of keywords and their total link strength (TLS)

Cluster	Keyword	Occurrences	Total Link Strength
Cluster 1 (Red)	Human	3	13
	Logistic regression analysis	3	11
	Article	3	11
	Logistic models	2	11
	Statistical model	2	11
	Humans	2	11
Cluster 2 (Green)	Logistic regression	2	9
	Firth logistic regression	3	4
	Complete separation	3	3
	Biomarkers	2	2
Cluster 3 (Blue)	Regression analysis	2	10
	Estimation method	2	10

Fig. 6 presents a network visualization of keyword co-occurrences related to complete separation in logistic regression using VOSviewer. The analysis includes 11 keywords that met the minimum threshold of two occurrences. The network map organizes the keywords into three distinct clusters based on their co-occurrence patterns and total link strength.

The thickness of the lines connecting the nodes represents the strength of the relationships between the keywords (Sarudin et al., 2024). For instance, "human" shows strong connections with other keywords within Cluster 1, indicating that research on logistic regression in human-related data is a major focus. In contrast, the relatively weaker links between Cluster 2 and Cluster 3 suggest that methodological developments such as the Firth correction are less connected to applications involving human related data. This indicates a potential research gap. Such gaps from this network map can be determined by examining the links between nodes. If two keywords lack a connection, it suggests a promising area for future exploration in complete separation in logistic regression research.

https://doi.org/10.24191/jcrinn.v10i2.535

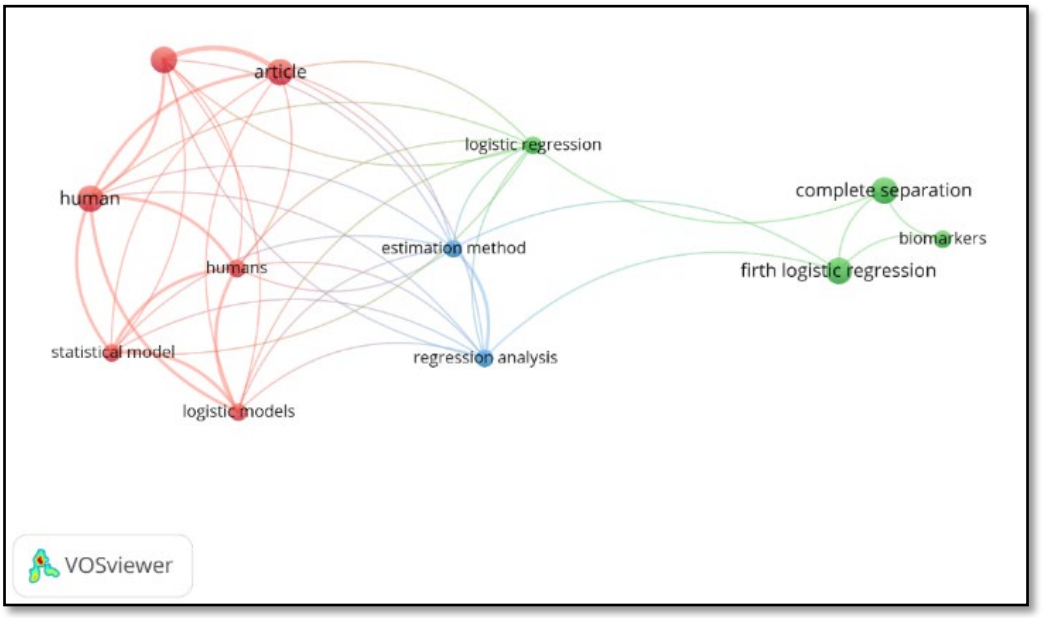

Fig. 6. Network node map of the exploration phase of complete separation in logistic regression research from 2012-2024.

Source: Online map: https://tinyurl.com/2brypyl9

5. CONCLUSION

This study presents the bibliometric analysis of research trends on complete separation in logistic regression using Firth penalized regression. By assessing nine selected articles published between 2012 and 2024, the study reveals a growing scholarly interest, particularly after 2019 with the United States and Malaysia emerging as the most productive contributors. Based on VOSviewer's co-authorship analysis, the United States also showed the highest collaboration connectivity, while Malaysia demonstrated strong regional collaboration despite its recent entry into the field. Influential articles primarily advanced methodological frameworks and applied solutions in health and transportation research. Keyword co-occurrence analysis identified main thematic clusters, such as human studies, statistical modelling, and estimation techniques, while also revealing underexplored areas. Although the small dataset reflects the niche and emerging nature of this research, the bibliometric analysis revealed significant trends in collaborations, publications, and thematic keyword evolution. The study acknowledges limitations such as the broad nature of some keywords such as 'human' and 'article', suggesting keyword normalization techniques for future research to improve thematic specificity. These findings lay a foundation for expanding interdisciplinary applications and addressing research gaps. Ultimately, this work not only enhances understanding of the academic landscape but also paves the way for methodological advancements and broader applications of Firth's approach in logistic regression.

6. ACKNOWLEDGEMENTS/FUNDING

The authors gratefully acknowledge the Faculty of Computer and Mathematical Sciences (FSKM) of Universiti Teknologi MARA (UiTM) at both the Tapah Campus and Shah Alam Campus for their institutional support and research facilities which made this work possible.

https://doi.org/10.24191/jcrinn.v10i2.535

7. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

8. AUTHORS' CONTRIBUTIONS

Nurul Husna Jamian: Study conception and design, Methodology development, Data analysis, Investigation, Original draft preparation. **Ahmad Zia Ul-Saufie Mohamad Japeri**: Supervision, Results validation, Manuscript review. **Mohammad Nasir Abdullah**: Research resources provision, Supervision, Manuscript review and editing.

9. REFERENCES

Abdullah, M. N., Wah, Y. B., Majeed, A. B. A., Zakaria, Y., & Shaadan, N. (2022a). Identification of blood-based multi-omics biomarkers for Alzheimer's disease using Firth's logistic regression. *Pertanika Journal of Science & Technology*, *30*(2), 1197–1218. https://doi.org/10.47836/pjst.30.2.19

Abdullah, M. N., Wah, Y. B., Zakaria, Y., Majeed, A. B. A., & Huat, O. S. (2022b). Discovering potential blood-based cytokine biomarkers for Alzheimer's disease using Firth logistic regression. *Epidemiology Biostatistics and Public Health*, *16*(4). https://doi.org/10.2427/13173

Alam, T. F., Rahman, M. S., & Bari, W. (2022). On estimation for accelerated failure time models with small or rare event survival data. *BMC Medical Research Methodology*, *22*, 169. https://doi.org/10.1186/s12874-022-01638-1

Allison, P. D. (2008). Convergence failures in logistic regression. In *SAS Global Forum 2008* (Vol. 360, No. 1, p. 11). http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Blondel, V., Guillaume, J. L., & Lambiotte, R. (2024). Fast unfolding of communities in large networks: 15 years later. *Journal of Statistical Mechanics Theory and Experiment*, *2024*, 10R001. https://doi.org/10.1088/1742-5468/ad6139

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Botes, M., & Fletcher, L. (2014, January). Comparing logistic regression methods for a sparse data set when complete separation is present. In *Annual Proceedings of the South African Statistical Association Conference* (Vol. 2014, No. con-1, pp. 1-8). South African Statistical Association (SASA).

Clark, R. G., Blanchard, W., Hui, F. K. C., Tian, R., & Woods, H. (2023). Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data. *Research Methods in Applied Linguistics*, *2*(1), 100044. https://doi.org/10.1016/j.rmal.2023.100044

Choi, L., Carroll, R. J., Beck, C., Mosley, J. D., Roden, D. M., Denny, J. C., & Van Driest, S. L. (2018). Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects. *Bioinformatics*, 34(17), 2988–2996. https://doi.org/10.1093/bioinformatics/bty306

D'Angelo, G., & Ran, D. (2024). Tutorial on Firth's logistic regression models for biomarkers in preclinical space. *Pharmaceutical Statistics, 24*(1), e2422. https://doi.org/10.1002/pst.2422

De Oliveira, O. J., Da Silva, F. F., Juliani, F., Barbosa, L. C. F. M., & Nunhes, T. V. (2019). Bibliometric method for mapping the state-of-the-art and identifying research gaps and trends in literature: An essential instrument to support the development of scientific projects. *IntechOpen.* https://doi.org/10.5772/intechopen.85856

Dobrescu, A., Nussbaumer-Streit, B., Klerings, I., Wagner, G., Persad, E., Sommer, I., Herkner, H., & Gartlehner, G. (2021). Restricting evidence syntheses of interventions to English-language publications is a viable methodological shortcut for most medical topics: A systematic review. *Journal of Clinical Epidemiology*, *137*, 209–217. https://doi.org/10.1016/j.jclinepi.2021.04.012

Donner, P. (2020). A validation of coauthorship credit models with empirical data from the contributions of PhD candidates. *Quantitative Science Studies*, *1*(2), 551-564. https://doi.org/10.1162/qss_a_00048

Fijorek, K., & Sokolowski, A. (2012). Separation-resistant and bias-reduced logistic regression: STATISTICA macro. *Journal of Statistical Software*, *Code Snippets, 47*(2), 1-12. https://doi.org/10.18637/jss.v047.c02

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Gilbert, N. (2022). *Logistic regression* (ebook). Taylor & Francis Group.

Gim, T. H. T., & Ko, J. (2016). Maximum likelihood and Firth logistic regression of the pedestrian route choice. *International Regional Science Review*, *40*(6), 616–637. https://doi.org/10.1177/0160017615626214

Harzing, A. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*, 341–349. https://doi.org/10.1007/s11192-019-03114-y

Harzing, A. W. (2023). *Publish or Perish user's manual*. Harzing.com. https://harzing.com/resources/publish-or-perish/manual

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409-2419.

Hess, A. S., & Hess, J. R. (2019). Logistic regression. *Transfusion*, *59*(7), 2197–2198. https://doi.org/10.1111/trf.15406

Ilmasari, D., Sahabudin, E., Riyadi, F. A., Abdullah, N., & Yuzir, A. (2022). Future trends and patterns in leachate biological treatment research from a bibliometric perspective. *Journal of Environmental Management*, *318*, 115594. https://doi.org/10.1016/j.jenvman.2022.115594

Iskandar, A., Azis, F., Dewi, R. D. C., Rusli, R., & Ahmar, A. S. (2020). Co-authorship visualization of research on COVID-19 from Web of science data using bibliometric analysis. *Library Philosophy and Practice (e-journal), 4528*, 1-9. https://digitalcommons.unl.edu/libphilprac/4528

Karabon, P. (2020, March). Rare events or non-convergence with a binary outcome? The power of Firth regression in PROC LOGISTIC. In *SAS Global Forum 2020*. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4654-2020.pdf

Kumar, R. (2025). Bibliometric analysis: Comprehensive insights into tools, techniques, applications, and

solutions for research excellence. *Spectrum of Engineering and Management Sciences*, *3*(1), 45–62. https://doi.org/10.31181/sems31202535k

Mansournia, M. A., Geroldinger, A., Greenland, S., & Heinze, G. (2017). Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology*, *187*(4), 864–870. https://doi.org/10.1093/aje/kwx299

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & López-Cózar, E. D. (2021). Google scholar, Microsoft academic, Scopus, Dimensions, Web of science, and open citations' coci: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*, 871–906. https://doi.org/10.1007/s11192-020-03690-4

Nathanson, B. H., & Higgins, T. L. (2008). An introduction to statistical methods used in binary outcome modeling. *Seminars in Cardiothoracic and Vascular Anesthesia*, *12*(3), 153–166. https://doi.org/10.1177/1089253208323415

Noor, H. M., & Asmael, N. M. (2023). A study on interstate freight mode choice between trucks and trains used to transport oil products: A case study of Iraq. *Transport Problems*, *18*(3), 29–40. https://doi.org/10.20858/tp.2023.18.3.03

Nussbaumer-Streit, B., Klerings, I., Dobrescu, A. I., Persad, E., Stevens, A., Garritty, C., Kamel, C., Affengruber, L., King, V. J., & Gartlehner, G. (2020). Excluding non-English publications from evidence-syntheses did not change conclusions: A meta-epidemiological study. *Journal of Clinical Epidemiology*, *118*, 42–54. https://doi.org/10.1016/j.jclinepi.2019.10.011

Olowe, K. J., Edoh, N. L., Zouo, S. J. C., & Olamijuwon, J. (2024). Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences*, *7*(2), 016–026. https://doi.org/10.53346/wjapls.2024.7.2.0039

Perianes-Rodriguez, A., Waltman, L., & Van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, *10*(4), 1178–1195. https://doi.org/10.1016/j.joi.2016.10.006

Pozsgai, G., Lövei, G. L., Vasseur, L., Gurr, G., Batáry, P., Korponai, J., Littlewood, N. A., Liu, J., Móra, A., Obrycki, J., Reynolds, O., Stockan, J. A., VanVolkenburg, H., Zhang, J., Zhou, W., & You, M. (2020). A comparative analysis reveals irreproducibility in searches of scientific literature. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.03.20.997783

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics In Medicine*, *36*(14), 2302–2317. https://doi.org/10.1002/sim.7273

Rahiminejad, S., Maurya, M. R., & Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics*, *20*, 212. https://doi.org/10.1186/s12859-019-2746-0

Rainey, C., & McCaskey, K. (2021). Estimating logit models with small samples. *Political Science Research and Methods*, *9*(3), 549–564. https://doi.org/10.1017/psrm.2021.9

Rojas-Flores, S., Ramirez-Asis, E., Delgado-Caramutti, J., Nazario-Naveda, R., Gallozzo-Cardenas, M., Diaz, F., & Delfin-Narcizo, D. (2023). An analysis of global trends from 1990 to 2022 of microbial fuel cells: A bibliometric analysis. *Sustainability*, *15*(4), 3651. https://doi.org/10.3390/su15043651

Sarudin, E. S., Ariffin, W. N. M., & Jamian, S. S. (2024). Mapping the landscape: A bibliometric analysis of staff scheduling optimization research trends and keywords evolution. *International Journal of Research and Innovation in Social Science*, *8*(8), 358–372. https://doi.org/10.47772/ijriss.2024.808029

Sarudin, E. S., Aziz, W. N. H. W. A., Saleh, S. S. M., & Arsad, R. (2023). An overview of bibliometric indices and keyword classification in shift scheduling. *International Journal of Academic Research in Economics and Management Sciences*, *12*(2), 289-302. https://doi.org/10.6007/ijarems/v12-i2/17317

Šinkovec, H., Geroldinger, A., & Heinze, G. (2019). Bring more data! —A good advice? Removing separation in logistic regression by increasing sample size. *International Journal of Environmental Research and Public Health*, *16*(23), 4658. https://doi.org/10.3390/ijerph16234658

Stolte, M., Herbrandt, S., & Ligges, U. (2024). A comprehensive review of bias reduction methods for logistic regression. *Statistics Surveys*, *18*, 139-162. https://doi.org/10.1214/24-ss148

Suhas, S., Manjunatha, N., Kumar, C. N., Benegal, V., Rao, G. N., Varghese, M., & Gururaj, G. (2023). Firth's penalized logistic regression: A superior approach for analysis of data from India's national mental health survey, 2016. *Indian Journal of Psychiatry*, *65*(12), 1208–1213. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_827_23

Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, *111*, 1053–1070. https://doi.org/10.1007/s11192-017-2300-7

Van Eck, N. J., & Waltman, L. (2023). *VOSviewer manual* (Version 1.6.19). Centre for Science and Technology Studies, Leiden University. https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.19.pdf

Walker, D. A., & Smith, T. J. (2019). Logistic regression under sparse data conditions. *Journal of Modern Applied Statistical Methods*, *18*(2), eP3372. https://doi.org/10.22237/jmasm/1604190660

Wun, M. K., Padula, A. M., Greer, R. M., & Leister, E. M. (2022). A review of 91 canine and feline red-bellied black snake (pseudechis porphyriacus) envenomation cases and lessons for improved management. *Australian Veterinary Journal*, *100*(7), 318–328. https://doi.org/10.1111/avj.13159

Yaman, A., Yoganingrum, A., Yaniasih, Y., & Riyanto, S. (2019). Tinjauan pustaka sistematis pada basis data pustaka digital: Tren riset, metodologi, dan coverage fields. *Jurnal Dokumentasi Dan Informasi*, *40*(1), 1-20. https://doi.org/10.14203/j.baca.v40i1.481

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, *13*(2), 157-170. https://doi.org/10.1093/pan/mpi009