Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/73/2024.18864

Visualization Analysis of International Polygenic Risk Scores Based on CiteSpace

Jia Du[1,2,*], Lingyin Zhang[1,3]

[1]Zhengzhou University, School of Public Health, Zhengzhou University, Zhengzhou, Henan Province, China

[2]zddujia@163.com
[3]2197016608@qq.com
*corresponding author

Abstract. Objective: Over the past two decades, polygenic risk scores (PRS) have made significant advancements in predicting the risk and aiding in the precision prevention of complex diseases. This study conducts a visualization analysis of literature in this field to explore its current development, hotspots, and trends, providing a reference for further PRS research. Methods: We retrieved relevant literature on polygenic scoring from the Web of Science database and performed a visual analysis using the CiteSpace software, examining publication volume, authors, and institutional collaboration. Results: A total of 8,433 articles were retrieved, and 4,499 were included after manual selection. Overall, publication volume showed an upward trend; the most prolific author was Brenner, Hermann, with 61 publications; Harvard University ranked as the top institution in terms of publication volume; and the United States was the leading country in terms of output. High-frequency keywords included genome-wide association. Conclusion: Collaboration among scholars and institutions is relatively limited, suggesting the need for strengthened cooperation. Relevant domestic studies are few, while international research hotspots mainly focus on schizophrenia, coronary heart disease, and Alzheimer's disease, providing a reference for PRS research in China.

Keywords: Polygenic Risk Score, Coronary Heart Disease, Schizophrenia, Visualization Analysis.

1. Introduction

Today, many common complex diseases have multiple recognized risk loci, as well as numerous genetic determinants that have an impact too small to be detected on a genome-wide scale with statistical significance [1]. A simple and intuitive way to transform genetic data into a predictor of disease susceptibility is to aggregate the effects of these genetic loci into a single index known as a genetic risk score, also referred to as a polygenic risk score (PRS) [1]. Currently, the volume of domestic literature on polygenic risk scores is relatively limited, and there is a lack of bibliometric analysis in this field that combines visual and textual elements from multiple perspectives. CiteSpace software can reveal the research status and development trends of a given field through visual mapping. This study utilizes CiteSpace software to perform bibliometric analysis and create visual maps, providing an overview of

DOI: 10.54254/2753-8818/73/2024.18864

the status, hotspots, and trends in international PRS research over the past 20 years, offering a reference for future research.

2. Materials and Methods

2.1. Literature Retrieval Strategy

This study uses the Web of Science database as its data source, selecting the core database and advanced search module. The search strategy is as follows: ((((((((((TS=("genetic risk score")) OR TS=("Genetic Risk Scores")) OR TS=("Risk Score, Genetic")) OR TS=("Risk Scores, Genetic")) OR TS=("Score, Genetic Risk")) OR TS=("Scores, Genetic Risk")) OR TS=("Polygenic Risk Score")) OR TS=("Polygenic Risk Scores")) OR TS=("Risk Score, Polygenic")) OR TS=("Risk Scores, Polygenic")) OR TS=("Score, Polygenic Risk")) OR TS=("Scores, Polygenic Risk")). The search period is from January 1, 2005, to July 19, 2024. Exclusion criteria include advertisements, newspapers, conference papers, theses, duplicate records, and irrelevant articles, yielding a total of 8,433 articles. Inclusion criteria are limited to English journal articles closely related to polygenic risk scores; two researchers independently screened the literature based on these criteria, resolving disagreements through discussion or with the guidance of a graduate advisor.

2.2. Research Tool

This study employs CiteSpace 6.2.R4 (Advanced) software as the research tool.

2.3. Data Processing Method

The annual publication volume data was analyzed and plotted as a line chart using Excel. The top five countries and institutions by publication volume were statistically analyzed using a three-line table created in Excel. Selected articles were exported in Reworks format and imported into CiteSpace 6.2.R4 (Advanced) for data processing. The processing period was set from January 2005 to December 2024, with a time slice of one year, and other parameters were set to default. Data on publication volume, authors, countries, institutions, and keywords were collected and visually analyzed, and visual maps were generated and interpreted.

3. Results

3.1. Analysis of Publication Volume

An analysis of the publication trends in polygenic risk score research from 2005 to 2024 reveals that output was low before 2009, followed by a gradual increase beginning in 2010. Starting in 2019, the number of publications on polygenic risk scores grew rapidly, showing an overall upward trend each year. Since the data for 2024 includes only publications up to July, it may not represent the full year, with the current peak in publications observed in 2023. See Figure 1.

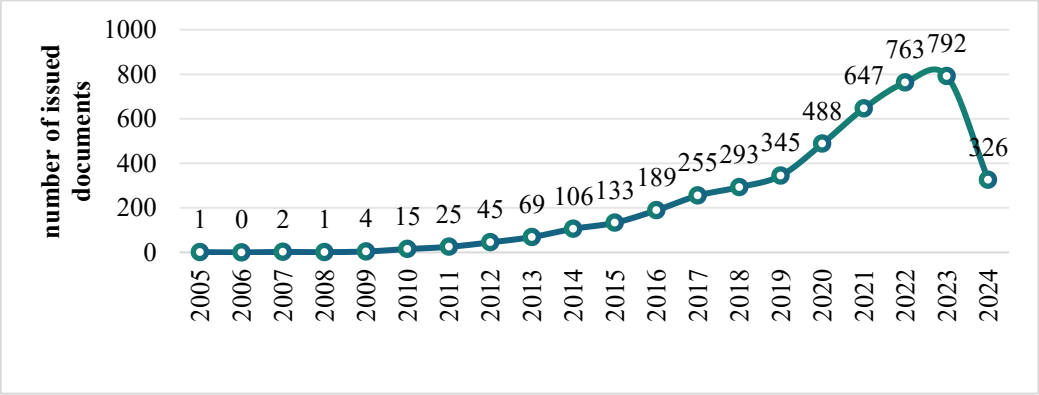

Figure 1. Annual Publication Volume on Polygenic Risk Scores

DOI: 10.54254/2753-8818/73/2024.18864

3.2. Analysis of Authors

By selecting "author" as the node type in CiteSpace, an analysis map was generated where each node (N) represents an author. The node size reflects publication volume, while the links (E) between nodes indicate collaboration among authors. A visual analysis of authors yielded 869 nodes and 3,775 links, with a density of 0.01. The top five authors by publication volume are Brenner, Hermann (61 publications); Kraft, Peter (56 publications); Andreassen, Ole A (50 publications); Zheng, Wei (50 publications); and Easton, Douglas F (48 publications). Easton, Douglas F collaborates closely with Brenner, Hermann, Kraft, Peter, and Zheng, Wei, while Andreassen, Ole A has limited connections with these four. See Figure 2.

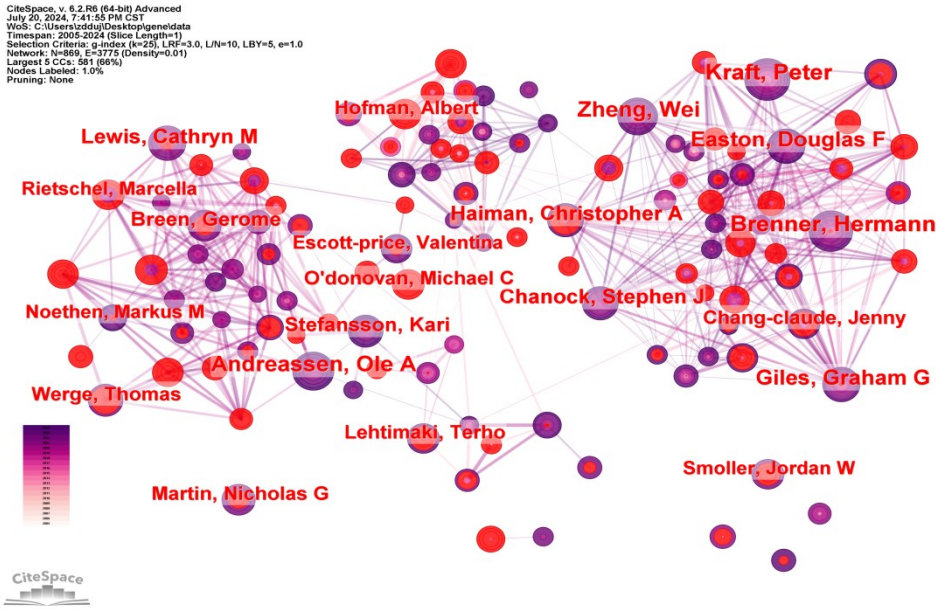

Figure 2. Visual Analysis of Authors with ≥10 Publications in Polygenic Risk Score Research

3.3. Analysis of Institutions

Using "institution" as the node type in CiteSpace, an analysis map was generated where each node (N) represents an institution. Node size reflects publication volume, and links (E) indicate collaboration between institutions. The analysis included 595 nodes and 6,042 links, with a density of 0.0342. Harvard University has the highest publication volume with 770 articles. Table 1 lists the top five institutions by publication volume. Due to the large number of institutions and complex collaborations, Figure 3 only displays institutions with ≥200 publications. Harvard Medical School is central, with strong collaborations with Harvard University, Broad Institute, Massachusetts Institute of Technology, University of California, Massachusetts General Hospital, Brigham and Women's Hospital, and Harvard T.H. Chan School of Public Health. University of London is central in another collaboration cluster, connecting with Helmholtz Association, King's College London, University College London, and University of Cambridge. The National Institutes of Health and Karolinska Institute have close ties to both clusters, while Icahn School of Medicine at Mount Sinai collaborates less. See Figure 3.

DOI: 10.54254/2753-8818/73/2024.18864

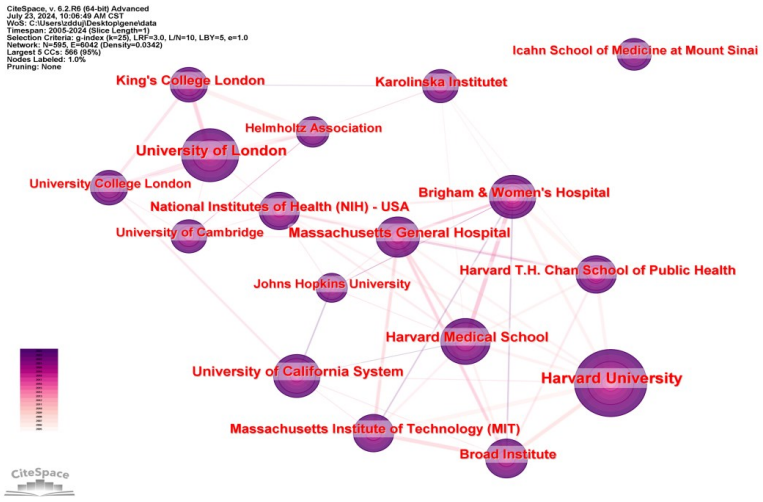

Figure 3. Visual Analysis of Institutions with ≥200 Publications in Polygenic Risk Score Research

Table 1. Top 5 Institutions by Publication Volume in Polygenic Risk Score Research

Rank	Institution	Number of Publications
1	Harvard University	770
2	University of London	532
3	Harvard Medical School	447
4	University of California System	364
5	Massachusetts General Hospital	343

3.4. Analysis of Publishing Countries

Using "country" as the node type in CiteSpace, an analysis map was created, where each node (N) represents a country, with larger nodes indicating higher publication volumes. Links (E) represent collaboration between countries. The visual analysis includes 115 nodes and 1,143 links, with a density of 0.1744. The United States has the highest number of publications, totaling 2,192, and maintains close collaborations with the United Kingdom, Netherlands, Canada, Germany, Australia, and Sweden. In contrast, China, Japan, and South Korea are less connected with this collaboration cluster. See Figure 4. Table 2 presents the top five countries by publication volume.

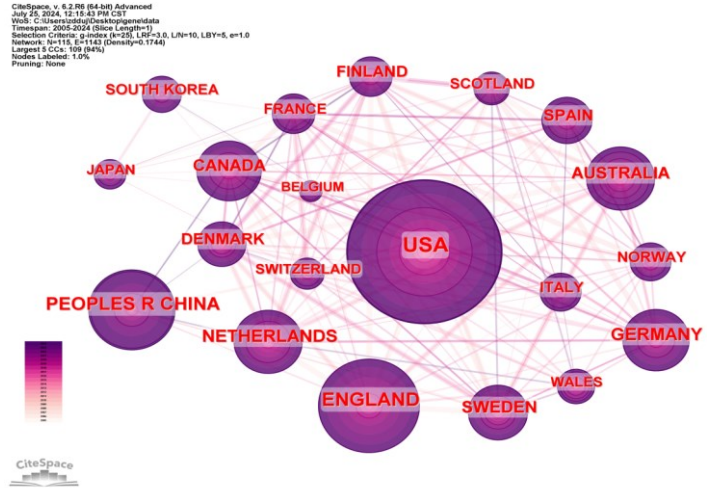

Figure 4. Visualization of Countries with ≥100 Publications in Polygenic Risk Score Research

DOI: 10.54254/2753-8818/73/2024.18864

Table 2. Top 5 Countries by Publication Volume in Polygenic Risk Score Research

Rank	Country	Number of Publications
1	USA	2192
2	ENGLAND	1047
3	PEOPLES R CHINA	789
4	NETHERLANDS	525
5	GERMANY	505

3.5. Analysis of High-Frequency Keywords

3.5.1. Keyword Co-Occurrence Analysis

Selecting "keyword" as the node type in CiteSpace produced an analysis map where each node (N) represents a keyword. The larger and more prominent the font, the higher the frequency of the keyword. Links (E) indicate relationships between keywords. The visual analysis includes 670 nodes and 7,000 links, with a density of 0.0312. See Figure 5. After excluding search terms related to "polygenic risk score," the most frequent keyword is "genome-wide association," appearing 1,147 times. Table 3 lists the top 10 keywords by frequency.

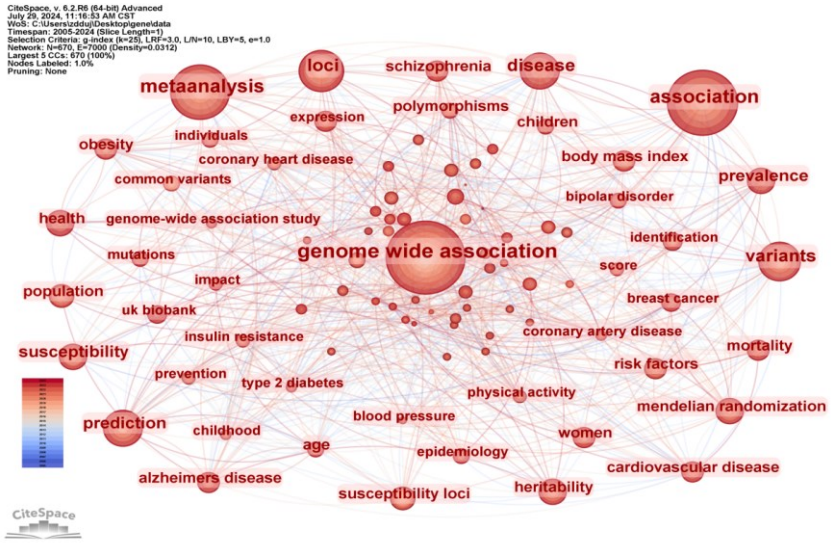

Figure 5. Co-Occurrence Map of Keywords with ≥50 Occurrences in Polygenic Risk Score Research

Table 3. Top 10 Keywords in Polygenic Risk Score Research

Rank	Keyword	Frequency
1	genome wide association	1147
2	association	915
3	metaanalysis	667
4	variants	494
5	loci	482
6	disease	407
7	prediction	361
8	prevalence	278
9	susceptibility	246
10	body mass index	228

DOI: 10.54254/2753-8818/73/2024.18864

3.5.2. Keyword Clustering Analysis

In CiteSpace, a clustering analysis of keywords was performed, yielding a modularity (Q) value of 0.3187 and a mean silhouette (S) score of 0.6391. With Q > 0.3 and S > 0.5, the clustering results are considered significant and reliable [2]. A total of seven clusters were identified: #0 Schizophrenia, #1 Coronary Artery Disease, #2 Obesity, #3 Breast Cancer, #4 Alzheimer's Disease, #5 Air Pollution, and #6 Single Nucleotide Polymorphism. See Figure 6.

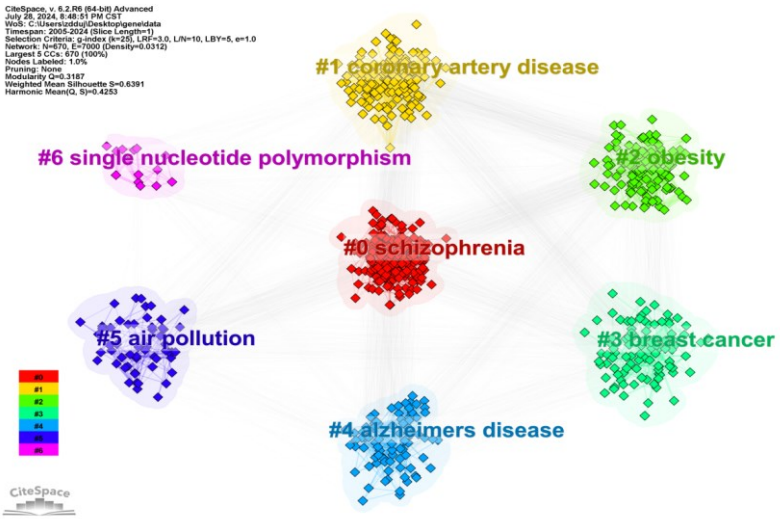

Figure 6. Keyword Clustering Map for Polygenic Risk Score Research

3.5.3. Keyword Burst Analysis

Using the Burstness function in CiteSpace, a burst analysis was conducted on the top 10 keywords. The results indicate that from 2005 to 2015, research primarily focused on cardiovascular diseases, with "coronary heart disease" showing the longest duration. From 2009 to 2018, the foundational aspects of polygenic risk scores, such as "genome-wide association," "polymorphism," "common variant," "loci," "replication," and "susceptibility loci," were key areas, with "genome-wide association" showing the highest burst strength. Since 2022, "polygenic risk score" has become the main research hotspot, with promising prospects for continued research.

Top 10 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2005 - 2024
coronary heart disease	2005	9.74	**2005**	2015	
cardiovascular disease	2007	10.75	**2007**	2014	
genome wide association	2009	22.84	**2009**	2016	
polymorphisms	2009	11.08	**2009**	2014	
common variants	2010	16.44	**2010**	2016	
locus	2010	8.96	**2010**	2017	
loci	2009	11.68	**2011**	2015	
replication	2012	11.13	**2012**	2017	
susceptibility loci	2011	12.13	**2015**	2018	
polygenic risk scores	2017	12.52	**2022**	2024	

Figure 7. Burst Analysis Map of the Top 10 Keywords in Polygenic Risk Score Research

DOI: 10.54254/2753-8818/73/2024.18864

4. Discussion

4.1. Core Author Analysis

The core author group, also known as prolific authors, consists of those with the highest publication volume and significant influence within a particular field [3]. Based on publication volume, core authors are determined using the formula proposed by the renowned scholar D. Price, $M = 0.749 \times \sqrt{N_{max}}$, where M is the minimum publication count for core authors, and Nmax is the publication count of the most prolific author. Between 2005 and 2024, the highest publication count by an author was 61, so Nmax=61. Substituting into the formula yields M≈5.85, making 6 the minimum publication count for core authors. Authors with publication counts between 6 and 61, totaling 271 individuals, represent 31.2% of the total authors, which does not reach half of the total authorship. According to Price's Law, core authors must constitute over half of all authors to be considered a core group, indicating a lack of a closely connected core author group in the field of polygenic risk scores [4].

Among them, Hermann Brenner, the most prolific author, has recently focused on polygenic risk scores for Alzheimer's disease [5] and colorectal cancer [6]. His research examines the association between polygenic risk scores, subjective cognitive decline (SCD), and the risk of dementia, aiming to assess whether these scores can enhance the established Cardiovascular Risk Factors, Aging, and Dementia (CAIDE) model, and how their predictive capabilities compare [5]. Other studies by core authors cover topics like glaucoma [7], venous thromboembolism [8], and schizophrenia [9], with Alzheimer's disease remaining at the forefront of research.

4.2. Analysis of Research Hotspots

Genome-wide association studies (GWAS) are designed to detect associations between genetic variations and traits in population samples. The primary aim of these studies is to enhance the understanding of disease biology to improve prevention and treatment. To date, GWAS relies on and utilizes linkage disequilibrium (LD), with the development of SNP arrays facilitating GWAS. Common SNP arrays vary in content but generally contain between 200,000 and 2,000,000 SNPs [10]. The advancement of GWAS has also propelled Mendelian randomization (MR) analysis, offering unique opportunities to identify causal relationships via MR. GWAS involves interrogating millions of single nucleotide polymorphisms (SNPs) to infer those associated with traits [11].

4.2.1. Schizophrenia

Schizophrenia is a chronic mental disorder with a heterogeneous genetic and neurobiological background, impacting early brain development and manifesting as a combination of psychotic symptoms (such as hallucinations, delusions, and disorganized thinking) along with motivational and cognitive impairments [12]. Genome-wide association studies (GWAS) have identified over a hundred related genetic loci for schizophrenia. While environmental factors play a role in the disorder, genetic effects contribute more significantly to its risk. Schizophrenia's genetic influence is not limited to single genes; rather, variations across chromosomal lengths provide a genetic risk score that offers a predictive value beyond randomness [13].

Antipsychotic medications are the primary treatment for schizophrenia; however, approximately 20-30% of patients show no response to these drugs and are classified as having treatment-resistant schizophrenia. Treatment-resistant schizophrenia is defined as a less than 20% reduction in positive symptoms after at least two trials of non-clozapine antipsychotics, with sufficient dosage and duration for each trial. GWAS has revealed the polygenic nature of treatment-resistant schizophrenia, and gene expression imputation has enabled the translation of GWAS findings into the construction of gene regulation and expression (GReX) risk scores (GReX-RS). GReX-RS for identified genes is constructed based on genotyped samples of psychotic patients to examine associations with clinical phenotypes, including clinical symptomatology, overall functioning, and cognitive performance, validating clinical predictors of treatment-resistant schizophrenia and providing a clinical explanation for its polygenic structure [14].

DOI: 10.54254/2753-8818/73/2024.18864

4.2.2. Coronary Heart Disease

Coronary artery disease (CAD), commonly referred to as coronary heart disease, is a chronic condition influenced by both genetic and environmental factors. Cardiovascular disease has surpassed cancer as the leading cause of death in China, underscoring the importance of early diagnosis and prevention for cardiovascular conditions [15]. For atherosclerotic cardiovascular disease, traditional clinical risk factors are effective in risk prediction; however, genetic factors may play a greater role in younger populations where clinical risk factors are less apparent, whereas in older populations, clinical risk factors tend to be more significant [16]. Genetic factors have a substantial impact on coronary artery disease, with around 60 gene loci strongly associated with the condition confirmed through common variant association studies [17]. GWAS has established coronary heart disease as one of the earliest applications of polygenic risk scores (PRS). PRS shows promise in identifying patients at higher risk of adverse outcomes, and incorporating PRS into clinical care could provide valuable guidance for risk stratification, helping to identify individuals who might benefit from early lifestyle interventions and medical treatment [18].

Numerous PRS have been developed and validated, consistently finding an increased risk of coronary artery disease among individuals in extreme percentiles compared to others. These findings have raised expectations for the clinical utility of PRS. However, the potential improvement in risk prediction by adding PRS to current guideline-based algorithms remains debated. Despite significant attention toward implementing PRS in routine clinical practice, several major challenges continue to limit its integration. Currently, no recommended strategy exists for combining these risk estimates [19].

4.2.3. Obesity

Obesity is a complex, multifactorial disease characterized by excessive body fat accumulation, which negatively impacts health [20]. In severe cases, it can lead to conditions such as gastrointestinal, liver, and cardiovascular diseases, and even potentially trigger diabetes or certain cancers. In recent years, obesity has become a global public health crisis and continues to worsen at an alarming rate. Although gut microbiota has been identified as a key contributor to the onset and progression of obesity and obesity-related diseases over the past decade [21], genetic factors remain a significant challenge for those affected by obesity. Identification of obesity-associated genetic regions is mainly conducted through genome-wide association studies, which compare the gene sequences of obese and non-obese individuals to identify specific genetic variations associated with this condition. To date, numerous SNPs (single nucleotide polymorphisms) linked to obesity have been identified. Given that each allele has a limited impact on obesity risk, a composite scoring system known as the Genetic Risk Score (GRS) has been developed. This system aggregates the cumulative effects of multiple obesity-related SNPs within an individual, integrating multiple SNPs related to BMI to generate a polygenic risk score through weighted calculation. This approach can predict the risk of genetic obesity, providing valuable insights for personalized prevention and treatment strategies. It is worth noting that most genome-wide association studies have been conducted in Western populations, primarily focusing on European demographics, with a notable lack of research on other populations [22].

4.2.4. Breast Cancer

Breast cancer is one of the most common cancers and a leading cause of cancer-related deaths among women worldwide. As a developing country, China's breast cancer survival rate is significantly lower than that of Western nations [23]. Furthermore, the currently effective mammography machines used for breast cancer detection are not fully suited to the physical characteristics of East Asian women, making detection more challenging. The BRCA1 and BRCA2 genes are widely recognized for their strong association with breast cancer, which makes genetic risk assessment particularly meaningful for breast cancer from a genetic perspective. Polygenic risk scores (PRS) are constructed using genotypes from multiple common risk loci, with various methods available for their development and application. These methods differ in the selection of genetic loci, the approach for determining the effect size of each locus, the statistical predictive models used for risk assessment, and the types of variables adjusted for.

Studies using a Bayesian regression framework have demonstrated that PRS stratification can effectively identify risk levels in women. Additionally, the importance of identifying genetic carriers (limited to three variants in CHEK2 and PALB2) and collecting family history data is reaffirmed [24].

5. Conclusion

This study used the CiteSpace software to analyze international research on polygenic risk scores, providing an initial visual representation of the research landscape in this field over the past 20 years. The results indicate a lack of collaboration among authors and institutions in related literature, suggesting a need for increased exchange and cooperation. A limitation of this study is its reliance on a single database, excluding other databases and lacking Chinese-language research. Future research could integrate domestic and international studies for a more comprehensive analysis.

References

[1] Igo, R. P., Jr., Kinzy, T. G., & Cooke Bailey, J. N. (2019). Genetic risk scores. Current Protocols in Human Genetics, 104(1), e95. https://doi.org/10.1002/cphg.95

[2] Gao, Z., Gao, L., Guo, X., et al. (2024). Visual analysis of hot spots in pre-diabetes nursing research at home and abroad based on CiteSpace. Evidence-Based Nursing, 10(14), 2544–2554.

[3] Xue, H., Si, X., & Zhang, F. (2023). Visual analysis of fall assessment scale-related research in China based on CiteSpace. Inner Mongolia Medical Journal, 55(12), 1480–1486. https://doi.org/10.16096/J.cnki.nmgyxzz.2023.55.12.016

[4] Geng, Y., & Liu, L. (2024). Knowledge map analysis of smart medical education under digital and technological empowerment. Chinese Medical Education Technology, 38(4), 416–422. https://doi.org/10.13566/j.cnki.cmet.cn61-1317/g4.202404005

[5] Trares, K., Stocker, H., Stevenson-Hoare, J., et al. (2024). Comparison of subjective cognitive decline and polygenic risk score in the prediction of all-cause dementia, Alzheimer's disease, and vascular dementia. Alzheimer's Research & Therapy, 16, 188. https://doi_org.libproxy.v.zzu.edu.cn/10.1186/s13195-024-01559-9

[6] Fu, R. M., Chen, X., Niedermaier, T., Seum, T., Hoffmeister, M., & Brenner, H. (2024). Excess weight, polygenic risk score, and findings of colorectal neoplasms at screening colonoscopy. The American Journal of Gastroenterology. https://doi.org/10.14309/ajg.0000000000002853

[7] de Vries, V. A., Hanyuda, A., Vergroesen, J. E., et al. (2024). The clinical utility of a glaucoma polygenic risk score in four population-based European-ancestry cohorts. Ophthalmology. https://doi.org/10.1016/j.ophtha.2024.08.005

[8] Jee, Y. H., Thibord, F., Dominguez, A., et al. (2024). Multi-ancestry polygenic risk scores for venous thromboembolism. medRxiv. https://doi.org/10.1101/2024.01.09.24300914

[9] Pentz, A. B., O'Connel, K. S., van Jole, O., et al. (2024). Mismatch negativity and polygenic risk scores for schizophrenia and bipolar disorder. Schizophrenia Research, 264, 314–326. https://doi.org/10.1016/j.schres.2024.01.013

[10] Visscher, P. M., Wray, N. R., Zhang, Q., et al. (2017). 10 years of GWAS discovery: Biology, function, and translation. American Journal of Human Genetics, 101(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

[11] Boehm, F. J., & Zhou, X. (2022). Statistical methods for Mendelian randomization in genome-wide association studies: A review. Computational and Structural Biotechnology Journal, 20, 2338–2351. https://doi.org/10.1016/j.csbj.2022.05.015

[12] Kahn, R., Sommer, I., Murray, R., et al. (2015). Schizophrenia. Nature Reviews Disease Primers, 1, 15067. https://doi.org/10.1038/nrdp.2015.67

[13] Toh, C., & Brody, J. P. (2021). A genetic risk score using human chromosomal-scale length variation can predict schizophrenia. Scientific Reports, 11, 18866. https://doi.org/10.1038/s41598-021-97983-0

[14] Gene expression imputation provides clinical and biological insights into treatment-resistant schizophrenia polygenic risk, Psychiatry Research, Volume 332, 2024, 115722, ISSN0165-1781, ttps://doi_org.libproxy.v.zzu.edu.cn/10.1016/j.psychres. 2024. 115722. (https://www_sciencedirect_com.libproxy.v.zzu.edu.cn/science/article/pii/S016517812400009X)

[15] Shi, Y., Hu, Y., Zhang, Y., et al. (2024). Impact of long-term ozone exposure on cardiovascular disease mortality in Ningxia region. Journal of Environmental and Health. http://kns.cnki.net/kcms/detail/12.1095.r.20240619.1035.002.html

[16] Tada, H., & Takamura, M. (2024). Assessment timings of polygenic risk score for atherosclerotic cardiovascular disease. Journal of Atherosclerosis and Thrombosis, 31(7), 1029–1030. https://doi.org/10.5551/jat.ED254

[17] Khera, A., & Kathiresan, S. (2017). Genetics of coronary artery disease: Discovery, biology, and clinical translation. Nature Reviews Genetics, 18, 331–344. https://doi.org/10.1038/nrg.2016.160

[18] Qin, M., Wu, Y., Fang, X., Pan, C., & Zhong, S. (2024). Polygenic risk score predicts all-cause death in East Asian patients with prior coronary artery disease. Frontiers in Cardiovascular Medicine, 11, 1296415. https://doi.org/10.3389/fcvm.2024.1296415

[19] Christoffersen, M., Stender, S., & Tybjærg-Hansen, A. (2024). Polygenic risk scores for cardiovascular risk prediction: Moving towards implementation into clinical practice? European Heart Journal, 45(20), 1853–1855. https://doi_org.libproxy.v.zzu.edu.cn/10.1093/eurheartj/ehae125

[20] Lin, X., & Li, H. (2021). Obesity: Epidemiology, pathophysiology, and therapeutics. Frontiers in Endocrinology, 12, 706978. https://doi.org/10.3389/fendo.2021.706978

[21] Geng, J., Ni, Q., Sun, W., Li, L., & Feng, X. (2022). The links between gut microbiota and obesity and obesity-related diseases. Biomedicine & Pharmacotherapy, 147, 112678, ISSN 0753-3322, https://doi_org.libproxy.v.zzu.edu.cn/10.1016/j.biopha.2022.112678.

[22] Chermon, D., & Birk, R. (2024). Deciphering the interplay between genetic risk scores and lifestyle factors on individual obesity predisposition. Nutrients, 16, 1296. https://doi_org.libproxy.v.zzu.edu.cn/10.3390/nu16091296

[23] Akram, M., Iqbal, M., & Daniyal, M. (2017). Awareness and current knowledge of breast cancer. Biological Research, 50, 33. https://doi_org.libproxy.v.zzu.edu.cn/10.1186/s40659-017-0140-9

[24] Maxwell, K. N., et al. (2024). Toward application of polygenic risk scores to both enhance and deintensify breast cancer screening. Journal of Clinical Oncology, 42, 1462–1465. https://doi.org/10.1200/JCO.24.00029