Knowledge Mining of Alzheimer's Disease Gene-Disease Associations: Postprint

Authors: Wang Xue, Wu Junwei, Chen Guanqun, Li Yanqiong, road

Date: 2023-04-01T16:15:56+00:00

Abstract

[Purpose/Significance] To conduct gene-disease association mining for Alzheimer's disease (AD) to capture potential research directions. [Methods/Process] An open knowledge discovery framework was constructed based on LBD theory. By integrating MeSH vocabulary, DisGeNET and other medical terminologies, and omics data, knowledge mining was performed on AD literature from PubMed. Association rules and algorithmic ranking methods were employed to screen strongly associated thematically co-occurring diseases with partial gene overlap and prioritize candidate genes, with validation conducted through temporal slicing and comparison with other LBD tools. [Results/Conclusion] Gene-disease identification was performed on 88,334 AD articles and matched with 2,120 AD-related genes. Association ranking was conducted on the identified 992 thematically co-occurring diseases and 11,899 candidate genes from an XYZ analysis perspective. Ten strongly associated diseases and 25 prioritized candidate genes were refined and elaborated with reference to literature reports. Mining potential associations among target diseases, co-occurring diseases, and genes through LBD can rapidly capture potential research directions, narrow the scope of gene sequencing, and provide important guidance for generating new research hypotheses.

Full Text

Knowledge Mining of Alzheimer's Disease Gene-Disease Associations

Wang Xue[1],[2], Wu Junwei[3], Chen Guanqun[4], Li Yanqiong[2], Ma Lu[1]

[1]School of Medical Humanities, Capital Medical University, Beijing 100069
[2]Library of Xuanwu Hospital, Capital Medical University, Beijing 100053
[3]Medical Information Section, Chinese PLA General Hospital, Beijing 100853

[4]Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing 100053

Abstract:
[**Purpose/Significance**] This study explores gene-disease associations in Alzheimer's disease (AD) to capture promising research directions. [**Method/Process**] Based on Literature-Based Discovery (LBD) theory, we constructed an open knowledge discovery framework. Using MeSH thesaurus, DisGeNET, and other medical terminologies and omics data, we performed knowledge mining on AD literature in PubMed. Association rules and algorithmic ranking methods were employed to screen strongly associated co-occurring diseases and priority candidate genes with partial gene overlap. Time slicing and other LBD tools were used for verification. [**Result/Conclusion**] From 88,334 AD articles, we identified gene-disease associations and matched them with 2,120 AD genes. Using an XYZ analysis perspective, we ranked 992 co-occurring diseases and 11,899 candidate genes. Ten strongly associated diseases and 25 priority candidate genes were refined, which can rapidly capture promising research directions, narrow the scope of gene sequencing, and provide important guidance for generating new research hypotheses.

Keywords: literature-based discovery; genomics; Alzheimer's disease; entity recognition; data mining; ranking algorithm; temporal analysis
Classification Numbers: G250; R745; R319.1
DOI: 10.13266/j.issn.0252-3116.2020.13.016

Introduction

Dementia has become one of the leading causes of mortality and disability among the elderly. Alzheimer's disease (AD), as the primary cause of dementia, represents one of the major global healthcare challenges of the 21st century. In 2015, the prevalence of AD and other dementias among people aged 60 and worldwide reached 5.2%, with the number of cases projected to double within 35 years. AD has a high disability rate, with patients in advanced stages losing independent living abilities and requiring continuous care, costing an estimated 1.09% of global GDP and imposing heavy burdens on families and society.

The complex pathogenesis has made breakthroughs difficult in this field over the past 40 years, with therapeutic drugs remaining primarily symptomatic without altering disease progression. Therefore, identifying AD risk factors and implementing early intervention or prevention are effective approaches to delay disease onset. Genetic factors represent the most definitive AD risk factor after age, and recent research has made significant progress. Although the amyloid-beta hypothesis has long dominated diagnostic and therapeutic development, genomic studies using linkage analysis, genome-wide association studies (GWAS), and massively parallel sequencing have revealed a series of biological processes contributing to AD and proposed new therapeutic targets, laying a foundation

for understanding the genetic causes of AD risk and explaining multifactorial complexity. While these results have limited impact on pathogenesis elucidation and treatment design, researchers remain optimistic about genetic studies clarifying new AD genes and the potential impact of genetic analysis on disease prevention.

Published scientific papers contain vast amounts of biomedical knowledge, including "established knowledge" validated by experiments and widely accepted, as well as "emerging knowledge" that has not yet gained widespread attention and lacks substantial research foundation. Although researchers tend to use established knowledge systems to explain questions, systematic analysis and validation of emerging knowledge are more conducive to transforming ideas into testable hypotheses, thereby stimulating in-depth exploration within disciplines and interdisciplinary collaboration. In 1986, D. Swanson proposed the Literature-Based Discovery (LBD) model, attempting to discover new, meaningful knowledge associations from existing literature through automated or semi-automated methods. LBD can be applied to drug side effect monitoring, new therapy research, and candidate disease gene identification.

LBD theory applied to AD knowledge discovery demonstrates rich layers, including analysis from perspectives of genes, protein molecules, metabolic products, and disease drugs to detect potential research directions in AD genetic variation, gene phenotype, and protein-cell physiology/pathophysiology. Although these efforts have yielded results, challenges remain, such as lack of external validation, limited data applicability, and difficulty in result interpretation.

This study constructs an open knowledge discovery architecture based on LBD theory (see Figure 1 [Figure 1: see original paper]), mines AD literature to identify associated diseases, combines gene-disease information from omics databases to infer AD potential candidate genes, and employs time slicing and other LBD tools for verification, aiming to provide references for clarifying AD pathogenesis and expanding diagnostic and treatment approaches.

1 Data Sources and Methods

1.1 Data Sources

PubMed is the most authoritative database in the international biomedical field, having indexed over 30.27 million articles. The vast biomedical knowledge contained in these articles constitutes a tremendous knowledge repository. The Medical Subject Headings (MeSH) thesaurus, compiled by the U.S. National Library of Medicine (NLM), is a hierarchical controlled vocabulary that can precisely and quickly reveal biomedical concepts in literature, ensuring effective retrieval of PubMed's massive collection. Term co-occurrence refers to keywords, title words, or subject headings representing literature topics appearing together in one article. Term co-occurrence relationships are important means for analyzing literature knowledge content and mining knowledge value, commonly used to predict associations between diseases and genes. Previous studies

have shown that co-occurrence analysis of MeSH terms in PubMed literature can successfully replicate D. Swanson's discoveries.

One of the greatest challenges in AD research is deciphering its underlying pathogenic mechanisms. The continuous development of molecular medicine enables biomedical research to effectively answer questions about gene-disease associations. Using text mining and multi-data source integration to automatically extract disease candidate genes from scientific literature and prioritize them is a strategy for obtaining molecular mechanism information. Currently, large amounts of omics information are integrated on public platforms, such as GeneCards, UniProtKB, and PharmGKB—integrated databases that annotate disease genetics based on genomics, proteomics, or pharmacogenomics. Comprehensive utilization of genomic, phenotypic, and environmental information resources can deepen researchers' understanding of disease mechanisms. Therefore, it is essential to cleverly integrate such gene-disease datasets and combine query terms, association terms, and database terminology through three-way co-occurrence relationships with rule-based pattern recognition algorithms to achieve gene prioritization, thereby providing ideas for next-generation sequencing directions.

After investigating 20 common bioinformatics databases, we selected six platforms based on disease coverage and data accessibility (see Table 1) to compile AD gene-disease association (GDA) data. Among them, DisGeNET demonstrates superior comprehensiveness and flexibility in identifying gene and disease vocabularies and supports disease annotation classification under terminologies including MeSH, UMLS, and ICD9-CM. Based on this, we exported all GDAs from DisGeNET as a gene annotation table for identifying MeSH co-occurring diseases, providing clues for potential knowledge association mining.

1.2 Research Methods

Specific steps are shown in the data processing flowchart (see Figure 2 [Figure 2: see original paper]):

1. Retrieved AD-related literature from PubMed using the search term "Alzheimer Disease"[Mesh], then performed deduplication and organization.
2. Obtained the MeSH thesaurus from NLM's FTP site (http://www.nlm.nih.gov/mesh/meshhome.html) and extracted disease category information from Category C (TreeNumbers: C).
3. Matched MeSH disease terms with AD literature subject terms to identify diseases co-occurring with AD.
4. Extracted gene-disease associations for co-occurring diseases using the complete disease GDA terminology from DisGeNET to obtain all disease-related genes.
5. Identified disease lists that completely overlapped with AD genes, partially contained AD genes, or contained no AD genes by matching against the

compiled AD gene set from six platforms. VBA programming was used for identification and matching, with processed data stored in an Access database.

2 Results and Analysis

2.1 Overall Results

As of July 31, 2019, the search retrieved 88,334 AD-related articles published since 1945. Using the MeSH Category C thesaurus (downloaded June 23, 2019) containing 11,648 categories/4,818 diseases, we performed disease entity recognition, identifying 166,946 instances/1,639 AD co-occurring diseases. Matching against 628,685 DisGeNET GDAs (downloaded August 2, 2019) identified 1,125 diseases, obtaining 151,710 associations/13,891 genes. We compiled 2,120 AD genes from six databases including ClinVar and MalaCards, and matched them with co-occurring disease genes, categorizing each disease by its relationship to AD genes: - 88 diseases had 135 associated genes completely overlapping with AD genes - 992 diseases had 13,891 associated genes partially containing AD genes, involving 1,992 AD genes and 11,899 non-AD genes - 45 diseases had 87 associated genes containing no AD genes

2.2 Correlation Analysis

LBD theory suggests that only novel links are meaningful. After pruning known concept pairs, remaining pairs (potential discoveries) are ranked to enable researchers to prioritize the most promising research directions. Therefore, this study focuses on co-occurring diseases with partial gene overlap with AD and their involved non-AD genes. Through association rules and algorithmic ranking of X (AD) → Y (partial gene overlap diseases) → Z (other disease genes), we achieve ranking of potential candidate genes to predict future research directions.

2.2.1 Co-occurring Disease Analysis Using the X→Y (confidence, support) association rule, we calculated values for AD and co-occurring disease connections, analyzing the top 10 results under descending XY and YZ associations (see Table 2 , Table 3):

confidence = Dx Dy
support = Dx Dy

Where Dx is the total number of AD articles; Dy is the number of articles for co-occurring diseases; Dx Dy is the number of articles where AD and co-occurring diseases appear together.

Amyloid Plaques and Vascular Dementia rank highly in both sorting methods. In terms of literature quantity (X→Y), Amyloid Plaques (4.06%, 3,501) has always been a major branch in AD research. In terms of gene overlap (Y→Z), its AD gene overlap percentage (87.18%, 204) also ranks at the top. As a hallmark

neuropathological change in AD, amyloid-beta (Aβ) and amyloid deposits play a crucial role in AD pathogenesis. The Aβ hypothesis has long posited that extracellular aggregation of Aβ deposits triggers neurodegenerative processes, leading to memory and cognitive loss and ultimately AD. However, clarifying how toxic substances are generated during Aβ formation and how they cause cellular dysfunction and death remains challenging. With advances in cryo-electron tomography (cryo-ET), researchers can conduct more in-depth studies on Aβ protein structure, plaque aggregation mechanisms, and their connection to AD, providing new ideas for diagnosis and drug development to delay or even prevent disease progression.

Vascular Dementia (VaD) (2.78%, 2,401) is the second most common dementia subtype, accounting for 5-10% of all cases. Caused by cerebrovascular and related lesions leading to cerebral blood perfusion 障碍, it results in local brain tissue damage and ultimately cognitive impairment or dementia. Many elderly dementia patients often exhibit both VaD and AD pathologies, and shared risk factors, Aβ deposition phenomena, and pathological factors such as nitric oxide-dependent mitochondrial abnormalities and cell division reveal commonalities in their pathogenic mechanisms, suggesting that cerebrovascular lesions and neurodegenerative processes may interact. Additionally, memory disorders, cerebral amyloid angiopathy, and tauopathies all rank highly, indicating their high research value in terms of research activity and genomics relevance.

2.2.2 Candidate Gene Analysis Since X→Z may have more than one intermediate Y, and X can reach Z through different Ys, we borrowed a heuristic ranking function for Z sorting (see Formula (3)) to filter strongly associated information. Using complete data through 2019 to extract AD candidate genes, we examined AD and potential gene relationships through PubMed and Entrez databases (see Table 4).

$$\mathrm{Rank(Zk)} = \Sigma(\mathrm{Sxyi} \times \mathrm{SyiZk})$$

Where Zk is the ranked candidate gene; Sxy and Syz are support values for X→Yi and Yi→Zk; m is the number of intermediate concepts Yi.

Data through 2019 extracted 11,899 candidate genes. Algorithmic screening identified 25 priority candidate genes with partial literature validation. Results show:

1. **SPP1** ranks first across all three sorting methods and both association value perspectives, showing exceptionally prominent associations. The SPP1 gene encodes Secreted Phosphoprotein 1/Osteopontin, expressed in brain and various tissues, participating in inflammation and anti-apoptosis processes, and functioning as a cell adhesion molecule and cytokine. In 2015, M. Shi et al. found that a cerebrospinal fluid 5-peptide panel including SPP1 protein showed significant specificity and sensitivity in distinguishing Parkinson's disease from AD. Subsequently, cerebrospinal fluid and urinary SPP1 protein were used as candidate diagnostic markers for

monitoring progression in MCI and preclinical AD. A. Rentsendorj elucidated that SPP1 protein can regulate macrophage-mediated Aβ clearance, suggesting that increased brain osteopontin correlates with decreased Aβ in AD models. W. Kamphuis and Z. Yin showed through transcriptional profiling of CD11c+ microglia and MHCII+ plaque-associated microglia in APP/PS1 mice that SPP1, as an upregulated gene, participates in cell differentiation and system development. In 2019, C. Frigerio's latest research on gene regulation of microglia to Aβ plaques found that Aβ plaque presence in AppNL-GF mice promoted redistribution of homeostatic microglia to activated microglia, with SPP1 as a tissue repair gene further distinguishing activated microglia subpopulations, helping reveal AD microglia pathological features. Since 2007, research reports on SPP1 as a gene regulatory product-protein molecular biomarker have been continuously produced, but studies directly exploring gene transcription, expression, and participation in AD pathogenesis remain limited. Based on previous research and our data, SPP1-AD related research warrants further in-depth investigation.

2. **PTHLH** ranks high under different association values. Other candidate genes show similar patterns (see Table 5), suggesting that although lacking direct literature support, these genes have been reported in association with other neurodegenerative diseases or nervous system diseases (NSD), allowing potential research directions to be mined through strong AD-NSD associations.

3 Evaluation

Evaluating LBD is challenging because captured discoveries have not been published in any field, making validity difficult to verify. However, understanding discovery reliability is crucial, primarily measured through gold standard sets and evaluation metrics. Researchers commonly use baseline comparison, classic discovery replication, time slicing, expert/user evaluation, or experimental validation, combined with quantitative metrics like information retrieval to test performance.

3.1 Time Slicing

3.1.1 Evaluation Protocol Time slicing is a primary LBD evaluation method. The dataset is divided into pre-discovery and post-discovery periods based on a cutoff date. Pre-period data is used to generate discoveries, and post-period data serves as a test set to develop gold standard sets for evaluation. Gold standard formulation varies depending on how association terms are evaluated and is not limited to post-period data extraction—expert opinions and patent trials can also create gold standard sets.

This study selects newly confirmed AD gene-disease associations in bioinformatics databases as the gold standard set, enabling more precise semantic extrac-

tion and relationship determination. Medicine requires time for discoveries to mature, and reasonable cutoff dates are crucial for hypothesis transformation, though date division lacks standards and is highly subjective. Based on AD literature development trends, we selected 2014-2015, when growth stabilized, as the time division range, designating December 31, 2014 as the cutoff point to extract pre-period AD candidate genes for validation against post-period newly confirmed AD gene sets. We employed information retrieval metrics including Precision (P), Recall (R), and F-Measure (F) for quantitative evaluation of both the full dataset and top-20 accuracy intervals.

3.1.2 Overall Evaluation Results As of December 31, 2014, literature data extracted 10,564 candidate genes. Comparison with post-2015 AD gene sets showed 380 successful predictions, with overall R = 0.8257 and P = 0.0359 (see Table 6). Using 11-point interpolated P-R curves, adjusted interpolated average precision (AiP) rose to 0.1260, indicating ranking positively impacts overall performance. Figure 3 [Figure 3: see original paper] shows P values decline significantly when R ranges from 0.0026 to 0.1003, suggesting predictions within the top 10% of R values (247 predictions, 38 successes) have strong accuracy. Weighted F-value curves show optimal P at R = 0.4011 when balancing or prioritizing comprehensive prediction (1,404 predictions, 152 successes), while F0.5 curves show optimal P at R = 0.2005 when prioritizing accuracy (533 predictions, 76 successes), indicating fewer predictions need to be reviewed when focusing on precision.

Literature searches revealed that some AD-associated genes predicted as failures before 2015 had already been published (e.g., ATP5PD, PMID: 23857120), but their associations were not recorded in DisGeNET and other databases before 2015, thus escaping identification in time slicing tests. This suggests scattered association gene information exists in PubMed beyond bioinformatics database records. Therefore, we established strict literature inclusion criteria, selecting studies on AD patients/animal models/related gene regulatory expression revealing positive/inverse associations as evidence resources to correct sorted results (see Figure 5 [Figure 5: see original paper]).

After correction, the top 20 successes could be obtained by reviewing 48 genes (see Table 7), with AP = 0.6933, a substantial improvement over pre-correction. The break-even point occurred at R = 0.65 (20 predictions, 13 successes), similar to pre-correction. F-value curves indicate high accuracy can be achieved by reviewing 22 predicted genes without prioritizing recall.

3.1.3 Interval Evaluation Results Since most researchers will not review all discoveries, evaluating the proportion of associations in top-k positions is important. Full-data P-R curves (Figure 3 [Figure 3: see original paper]) reveal higher success probabilities for top-ranked discoveries. Intercepting the top 20 successfully predicted candidate genes as a detection interval for re-evaluation, we searched PubMed to supplement supporting literature for false-positive genes

strongly associated with AD.

3.2 Other Evaluation Methods

Other LBD models often struggle to 回溯 historical data. This study uses real-time output combined with literature estimation for horizontal comparison with BITOLA, another XYZ theory and association rule-based system. Both use AD as the X concept, selecting the top 50 diseases under descending XY association values as Y concepts to list Z candidate gene lists. BITOLA and our approach identified 4,211 and 5,252 candidate genes respectively. Table 8 shows vastly different rankings for top-20 predictions, though genes ranked high in one system often appear later in the other. BITOLA's predictions contained fewer candidate genes with close supporting literature (indicating gene regulatory expression or biomarkers), while some of its predicted genes like TIMP1 and EPO were already marked as AD-associated genes in DisGeNET. Current data cannot fully demonstrate performance differences, requiring expanded range testing, multiple topic trials, or other gold standard sets to 完善 evaluation.

Data mining and text mining research findings partially corroborate our predictions. F. Yao identified differential urinary SPP1 protein expression in AD patients through iTRAQ experiments, reporting it as a urinary biomarker for early AD. Y. Cruz-Rivera et al. used microarray datasets combined with Traveling Salesman Problem path analysis to identify differential expression between AD patients and controls, finding FTL in the most relevant cycle as a potential AD biomarker. Similar conclusions from different studies provide evidence for prediction validity.

Discussion

Early knowledge discovery research focused on databases, but with emerging technologies and application models, emphasis has shifted to knowledge extraction from unstructured data (text). As an important branch of knowledge discovery, LBD research in bioinformatics mining is increasingly extensive, with continuous technological refinement including: (1) Data types: applying LBD to patents, case reports, and other non-paper types; (2) Analysis units: using controlled vocabularies like UMLS, MeSH, and Entrez Gene to facilitate cross-disciplinary discovery; (3) Processing workflows: proposing automated techniques based on concepts, relationships, graphs, or link prediction beyond manual detection; (4) Filtering mechanisms: eliminating noise associations at article, paragraph, and sentence levels before term-level filtering to narrow discovery scope; (5) Ranking technologies: using machine learning models for potential association ranking beyond conventional statistical sorting; (6) Result presentation: employing semantic typing, graphical visualization, matrix visualization, or discovery pathway techniques beyond association lists; (7) Discovery evaluation: adopting combined quantitative and qualitative assessment methods. LBD has been practiced in new drug development, drug repurposing, and

adverse event prediction, but still lacks the ability to mimic true concept connection formation, requiring integration of logic and optimized reasoning mechanisms to better understand complex associations. Moreover, LBD discoveries are exploratory hypotheses based on existing text, ultimately requiring end-user acceptance, making expert evaluation and user interaction studies essential for validity verification.

Building on previous LBD research, this study attempts to integrate multi-source omics data to better meet association discovery needs and uses it as a systematic evaluation standard to ensure gold standard set accuracy. Combining GDA data for bidirectional analysis of AD co-occurring diseases and involved candidate genes with multiple sorting to refine strongly associated diseases and priority candidate genes facilitates precise prediction scope for potential genes, effectively guiding research directions and saving time and costs. Multiple evaluation through time slicing, literature correction, and horizontal LBD system comparison shows that reviewing the top 20-22 AD candidate genes predicted by this study achieves optimal precision, demonstrating performance and effectiveness.

Limited by database and thesaurus scope, this study only performed entity recognition on diseases indexed by subject terms in Medline literature. Expanding literature scope to the full PubMed database or other databases while applying more medical terminologies like Emtree and UMLS for mapping could enhance gene-co-occurring disease associations. Our identification rules are primarily based on subject term co-occurrence, making it difficult to provide strong evidence for causal pathogenic mechanisms, though this does not affect disease-gene association extraction. This study focuses on genes and does not explore other biomedical concepts (proteins, cells, metabolic products), which will be considered in future research using more relationship types to strengthen discovery connections, construct heterogeneous networks, and combine ontologies or visualization techniques to extend AD knowledge discovery research. Additionally, considering common LBD issues like external validation and real-world data applicability, future collaboration with clinical and basic research teams will deepen related studies.

Rapid bioinformatics development has made important contributions to neuroscience, and linking genotype to phenotype for new association discovery remains a major challenge in etiological research for neurodegenerative diseases like AD. This study aims to rapidly capture more promising research directions through AD knowledge mining, further narrow gene sequencing scope, assist researchers in focusing on more valuable targets, provide important guidance for new research hypothesis generation, and offer crucial references for clarifying AD pathogenesis and expanding diagnostic and treatment approaches.

References

[1] Chinese Dementia and Cognitive Impairment Diagnosis and Treatment

Guidelines Writing Group, Chinese Medical Doctor Association Neurology Branch Cognitive Impairment Disease Professional Committee. 2018 Chinese dementia and cognitive impairment diagnosis and treatment guidelines (VII): Risk factors for Alzheimer's disease and their intervention[J]. Chinese Medical Journal, 2018, 98(19): 1461-1466.
[2] Scheltens P, Blennow K, Breteler MM, et al. Alzheimer's disease[J]. Lancet, 2016, 388(10043): 505-517.
[3] Prince M, Wimo A, Guerchet M, et al. World Alzheimer report 2015-the global impact of dementia[M]. London: Alzheimer's Disease International, 2015.
[4] Taylor CA, Greenlund SF, McGuire LC, et al. Deaths from Alzheimer's disease-United States, 1999-2014[J]. MMWR Morbidity and Mortality Weekly Report, 2017, 66(20): 521-527.
[5] Jia J, Wei C, Chen S, et al. The cost of Alzheimer's disease in China and re-estimation of costs worldwide[J]. Alzheimer's & Dementia, 2018, 14(4): 483-491.
[6] Patterson C. World Alzheimer report 2018-the state of the art of dementia research: new frontiers[M]. London: Alzheimer's Disease International, 2018.
[7] Verheijen J, Sleegers K. Understanding Alzheimer disease at the interface between genetics and transcriptomics[J]. Trends in Genetics, 2018, 34(6): 434-447.
[8] Van Cauwenberghe C, Van Broeckhoven C, Sleegers K. The genetic landscape of Alzheimer disease: clinical implications and perspectives[J]. Genetics in Medicine, 2016, 18(5): 421-430.
[9] Malhotra A, Younesi E, Gurulingappa H, et al. 'Hypothesis finder': a strategy for the detection of speculative statements in scientific text[J]. PLoS Computational Biology, 2013, 9(7): e1003117.
[10] Henry S. Indirect relatedness evaluation and visualization for literature-based discovery[D]. Virginia: Virginia Commonwealth University, 2019.
[11] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge[J]. Perspectives in Biology and Medicine, 1986, 30(1): 7-18.
[12] Henry S, McInnes BT. Literature-based discovery: models, methods, and trends[J]. Journal of Biomedical Informatics, 2017, 74: 20-32.
[13] Cohen T, Schvaneveldt RW. The trajectory of scientific discovery: concept co-occurrence and converging semantic distance[J]. Studies in Health Technology and Informatics, 2010, 160(1): 661-665.
[14] Hristovski D, Rindflesch T, Peterlin B. Using literature-based discovery to identify novel therapeutic approaches[J]. Cardiovascular & Hematological Agents in Medicinal Chemistry, 2013, 11(1): 14-24.
[15] Kim YH, Beak SH, Charidimou A, et al. Discovering new genes in the pathways of common sporadic neurodegenerative diseases: a bioinformatics approach[J]. Journal of Alzheimer's Disease, 2016, 51(1): 293-312.
[16] Kawalia A, Raschka T, Naz M, et al. Analytical strategy to prioritize Alzheimer's disease candidate genes using gene regulatory networks and public expression data[J]. Journal of Alzheimer's Disease, 2017, 59(4): 1237-1254.
[17] Gubiani D, Fabbretti E, Cestnik B, et al. Outlier-based literature explo-

ration for cross-domain linking of Alzheimer's disease and gut microbiota[J]. Expert Systems with Applications, 2012, 10: 217.
[18] Greco I, Day N, Riddoch-Contreras J, et al. Alzheimer disease biomarker discovery using in silico literature mining[J]. Genome Medicine, 2017, 9(1): 1-12.
[19] Malhotra A, Younesi E, Bagewadi S, et al. Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease[J]. BMC Genomics, 2014, 15(Suppl 1): S9.
[20] Smalheiser NR, Swanson DR. Linking estrogen to Alzheimer's disease: an informatics approach[J]. Neurology, 1996, 47(3): 809-810.
[21] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery[J]. Journal of Biomedical Informatics, 2006, 39(6): 600-611.
[22] Smalheiser NR, Swanson DR. Indomethacin and Alzheimer's disease[J]. Neurology, 1996, 46(2): 583.
[23] Li J, Zhu XY, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts[J]. PLoS Computational Biology, 2009, 5(7): e1000450.
[24] Chen R, Lin HF, Yang ZH. Passage retrieval based hidden knowledge discovery from biomedical literature[J]. Expert Systems with Applications, 2011, 38(8): 9958-9964.
[25] Zhang R, Simon G, Yu F. Advancing Alzheimer's research: a review of big data promises[J]. International Journal of Medical Informatics, 2017, 106: 48-56.
[26] Raja K, Patrick M, Gao Y, et al. A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries[J]. International Journal of Genomics, 2017, 2017: 6213474.
[27] Liu Q, Sun CP, Wang Q, et al. The role of PubMed database inclusion in enhancing international influence of medical journals[J]. Chinese Journal of Scientific and Technical Periodicals, 2015, 26(12): 1344-1347.
[28] Liu JH, Yu JR, Miao YG. Research on semi-automatic construction method of disease ontology based on MeSH thesaurus and co-word analysis[J]. Modern Information, 2009, 29(3): 208-211.
[29] Zhang YQ, Leng FH. Theoretical foundation research on knowledge discovery from non-related literature[J]. Journal of Library Science in China, 2009, 35(4): 25-30.
[30] Ruan GC. Research on topic model and text knowledge discovery applications[M]. Shanghai: East China Normal University Press, 2018.
[31] Sehgal A, Qiu X, Srinivasan P. Analyzing LBD methods using a general framework[C]//Bruza P, Weeber M. Literature-based discovery. Berlin: Springer, 2008: 75-100.
[32] Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering[J]. Scientometrics, 2003, 56(1): 111-135.
[33] Stegmann J, Grohmann G. Advanced information retrieval for hypothesis generation[C]//Society for Information Science. International workshop on webometrics, informetrics and scientometrics. Roorkee: Central Library,

Indian Institute of Technology, 2004: 334-346.
[34] Ono T, Kuhara S. A novel method for gathering and prioritizing disease candidate genes based on construction of a set of disease-related MeSH(R) terms[J]. BMC Bioinformatics, 2014, 15: 179.
[35] Piñero J, Queralt-Rosinach N, Bravo A, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes[J]. Database, 2015: bav028.
[36] Rappaport N, Fishilevich S, Nudel R, et al. Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect[J]. Biomedical Engineering Online, 2017, 16(S1): 72.
[37] Shui QY. Big data analysis for bioinformatics and biomedical discoveries[M]. Portland: CRC Press, 2016.
[38] Fayyad UM. Advances in knowledge discovery and data mining[M]. California: AAAI Press, 1996.
[39] Hristovski D, Peterlin B, Mitchell JA, et al. Using literature-based discovery to identify disease candidate genes[J]. International Journal of Medical Informatics, 2005, 74(2-4): 289-298.
[40] Maki S. The amyloid hypothesis on trial[J]. Nature, 2018, 559(7715): S4-S7.
[41] Iadanza MG, Jackson MP, Hewitt EW, et al. A new era for understanding amyloid structures and disease[J]. Nature Reviews Molecular Cell Biology, 2018, 19(12): 755-773.
[42] Alzheimer's Association. Vascular dementia[EB/OL]. [2019-12-24]. https://www.alz.org/alzheimers-dementia/what-is-dementia/types-of-dementia/vascular-dementia.
[43] Wu JH. Comparison of pathological mechanisms and related clinical research between Alzheimer's disease and vascular dementia[J]. Zhejiang Medical Journal, 2019, 41(11): 1227-1231.
[44] Ashraf GM, Chibber S, Mohammad I, et al. Recent updates on the association between Alzheimer's disease and vascular dementia[J]. Medicinal Chemistry, 2016, 12(3): 226-237.
[45] Chinese Medical Doctor Association Neurology Branch Cognitive Impairment Professional Committee, Chinese Vascular Cognitive Impairment Diagnosis and Treatment Guidelines Writing Group. 2019 Chinese guidelines for diagnosis and treatment of vascular cognitive impairment[J]. Chinese Medical Journal, 2019, 99(35): 2737-2744.
[46] Wung JK, Perry G, Kowalski A, et al. Increased expression of the remodeling and tumorigenic associated factor osteopontin in pyramidal neurons of the Alzheimer's disease brain[J]. Current Alzheimer Research, 2007, 4(1): 67-72.
[47] Shi M, Movius J, Dator R, et al. Cerebrospinal fluid peptides as potential Parkinson disease biomarkers: a staged pipeline for discovery and validation[J]. Molecular & Cellular Proteomics, 2015, 14(3): 544-555.
[48] Begcevic I, Brinc D, Brown M, et al. Brain-related proteins as potential CSF biomarkers of Alzheimer's disease: a targeted mass spectrometry approach[J]. Journal of Proteomics, 2018, 182: 12-20.
[49] Yao F, Hong X, Li S, et al. Urine-based biomarkers for Alzheimer's

disease[J]. Journal of Alzheimer's Disease, 2018, 65(2): 421-431.
[50] Rentsendorj A, Sheyn J, Fuchs DT, et al. A novel role for osteopontin in macrophage-mediated amyloid-β clearance in Alzheimer's disease models[J]. Journal of Neuroinflammation, 2018, 15(1): 1-15.
[51] Kamphuis W, Kooijman L, Schettens S, et al. Transcriptional profiling of CD11c-positive microglia accumulating around amyloid plaques in a mouse model for Alzheimer's disease[J]. Biochimica et Biophysica Acta, 2016, 1862(10): 1847-1860.
[52] Yin Z, Raj D, Saiepour N, et al. Immune hyperreactivity of Aβ plaque-associated microglia in Alzheimer's disease[J]. Neurobiology of Aging, 2017, 55: 115-122.
[53] Sala Frigerio C, Wolfs L, Fattorelli N, et al. The major risk factors for Alzheimer's disease: age, sex, and genes modulate the microglia response to Aβ plaques[J]. Cell Reports, 2019, 27(4): 1293-1306.
[54] Yetisgen-Yildiz M, Pratt W. Evaluation of literature-based discovery systems[C]//Bruza P, Weeber M. Literature-based discovery. Berlin: Springer, 2008: 101-113.
[55] Thilakaratne M, Falkner K, Atapattu T. A systematic review on literature-based discovery workflow[J]. PeerJ Computer Science, 2019, 5: e235.
[56] Hristovski D, Stare J, Peterlin B, et al. Supporting discovery in medicine by association rule mining in Medline and UMLS[J]. Studies in Health Technology and Informatics, 2001, 84(2): 1344-1348.
[57] Henry S, McInnes BT. Indirect association and ranking hypotheses for literature-based discovery[J]. BMC Bioinformatics, 2019, 20(1): 425.
[58] Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems[J]. Journal of Biomedical Informatics, 2009, 42(4): 633-643.
[59] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval[J]. Journal of the American Medical Informatics Association, 2005, 12(3): 296-298.
[60] Carterette BA, Voorhees EM. Overview of information retrieval evaluation[C]//Lupu M, Mayer K, Tait J, et al. Current challenges in patent information retrieval. Berlin: Springer, 2011: 69-85.
[61] Cruz-Rivera YE, Perez-Morales JL, Santiago YM, et al. A selection of important genes and their correlated behavior in Alzheimer's disease[J]. Journal of Alzheimer's Disease, 2018, 65(1): 193-205.
[62] Cifuentes RA, Murillo-Rojas J. Alzheimer's disease and HLA-A2: linking neurodegenerative to immune processes through an in silico approach[J]. BioMed Research International, 2014: 791238.
[63] Gopalakrishnan V, Jha K, Jin W, et al. A survey on literature-based discovery approaches in biomedical domain[J]. Journal of Biomedical Informatics, 2019, 93: 103141.
[64] Hofmann-Apitius M, Ball G, Gebel S, et al. Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders[J]. International Journal of Molecular Sciences, 2015, 16(12): 29179-29206.

[65] Pletscher-Frankild S, Palleja A, Tsafou K, et al. DISEASES: text mining and data integration of disease-gene associations[J]. Methods, 2015, 74: 83-89.

Author Contributions

Wang Xue: Developed research framework, designed methodology, collected materials, organized data, wrote and revised manuscript.
Wu Junwei: Participated in methodology design, programming and data processing, contributed to manuscript.
Chen Guanqun: Participated in methodology design, provided academic consultation, contributed to manuscript.
Li Yanqiong: Assisted in framework development, participated in manuscript revision.
Ma Lu: Proposed research idea, participated in manuscript revision.

Note: The original manuscript included journal policy statements regarding academic integrity and anti-plagiarism commitments. These have been omitted from the translation as they represent standard publication boilerplate rather than research content.

Note: Figure translations are in progress. See original paper for figures.