Journal of Advanced Research 40 (2022) 223–231

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Article

NDDRF: A risk factor knowledgebase for personalized prevention of neurodegenerative diseases

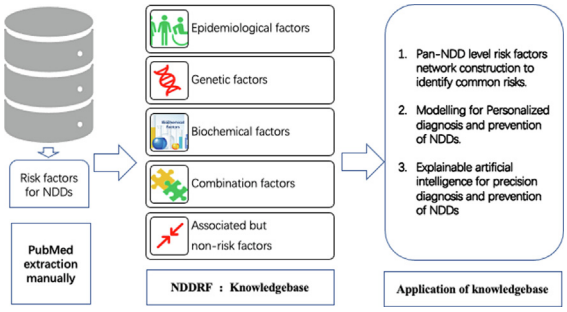
Cheng Bi [a,b], Shengrong Zhou [b], Xingyun Liu [a,b], Yu Zhu [a,c], Jia Yu [a,d], Xueli Zhang [b], Manhong Shi [b], Rongrong Wu [a,b], Hongxin He [b], Chaoying Zhan [a,b], Yuxin Lin [b], Bairong Shen [a,*]

[a] Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610212, Sichuan, China
[b] Center for Systems Biology, Soochow University, Suzhou 215006, Jiangsu, China
[c] Department of Bioinformatics, School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, Jiangsu, China
[d] School of Clinical Medicine, Soochow University, Suzhou 215123, Jiangsu, China

HIGHLIGHTS

- A risk factor knowledgebase (NDDRF) is built for neurodegenerative diseases (NDDs).
- NDDRF collects the risk factors associated with diagnosis and prevention of NDDs.
- NDDRF is helpful to the systematic understanding of the heterogeneous NDDs
- NDDRF provides knowledge for personalized diagnosis and prevention of NDDs.
- NDDRF can be used to the future explainable artificial intelligent modeling.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 January 2021
Revised 1 June 2021
Accepted 15 June 2021
Available online 20 June 2021

Keywords:
Neurodegenerative diseases
Risk factor
Protective factor
Knowledge base
Diagnosis and prevention

ABSTRACT

Introduction: Neurodegenerative diseases (NDDs) are a series of chronic diseases, which are associated with progressive loss of neuronal structure or function. The complex etiologies of the NDDs remain unclear, thus the prevention and early diagnosis of NDDs are critical to reducing the mortality and morbidity of these diseases.
Objectives: To provide a systematic understanding of the heterogeneity of the risk factors associated with different NDDs (pan-neurodegenerative diseases or pan-NDDs), the knowledgebase is established to facilitate the personalized and knowledge-guided diagnosis, prevention and prediction of NDDs.
Methods: Before data collection, the medical, life science and informatics experts as well as the potential users of the database were consulted and discussed for the scope of data and the classification of risk factors. The PubMed database was used as the resource of the data and knowledge extraction. Risk factors of NDDs were manually collected from literature published between 1975 and 2020.
Results: The comprehensive risk factors database for NDDs (NDDRF) was established including 998 single or combined risk factors, 2293 records and 1071 articles relevant to the 14 most common NDDs. The single risk factors are classified into 3 categories, i.e. epidemiological factors (469), genetic factors (324) and

Peer review under responsibility of Cairo University.
* Corresponding author.
E-mail address: bairong.shen@scu.edu.cn (B. Shen).

https://doi.org/10.1016/j.jare.2021.06.015

biochemical factors (153). Among all the factors, 179 factors are positive and protective, while 880 factors have negative influence for NDDs. The knowledgebase is available at http://sysbio.org.cn/NDDRF/.
Conclusion: NDDRF provides the structured information and knowledge resource on risk factors of NDDs. It could benefit the future systematic and personalized investigation of pan-NDDs genesis and progression. Meanwhile it may be used for the future explainable artificial intelligence modeling for smart diagnosis and prevention of NDDs.

Introduction

Neurodegenerative diseases (NDDs) are a series of chronic diseases that lead to progressive loss of neuronal structure or function [1–4]. According to World Health Organization' s forecasts, NDDs will overtake cancer and become the second leading cause of death, ranked after cardiovascular disease [5]. Distinct pathophysiological mechanisms underlie the different NDDs. Among them, Parkinson's disease (PD), which is the second most common NDDs in the elderly, begins with a prodromal period of several years and can last for more than 20 years after the onset of typical motor symptoms [6]. PD and amyotrophic lateral sclerosis (ALS) mainly cause motor symptoms whereas others, such as Alzheimer's disease (AD) mainly affect non-motor cognitive functions [7]. In 2018, about 50 million people worldwide suffered from dementia, it was predicted this number will increase to 152 million by 2050 [8].

There have been shown many similarities between different NDDs at the sub-cellular level, including atypical protein assembly and induced cell death. These similarities offer a hope, that is a single therapeutic strategy may be effective for the treatment of different NDDs [9,10]. Therefore, the pan-NDDs level information will benefit the systematic comparison of the profiles about different NDDs [11,12]. We previously established a database (NDDVD), collected the mutations associated with NDDs (http://www.sysbio.org.cn/NDDVD) and studied their similarity of mutation patterns [13]. With the accumulation of data from multiple NDDs omics measurement, we could identify the key players from the NDDs big data, such as biomarkers [14–16], drugs or drug targets [17] and risk factors for the prevention and prediction of NDDs [18].

Although the identification of risk factors for a particular disease is one of the most important areas of medical inquiry, such research is often aimed at discovering new causes of disease [19]. In NDDs, a comprehensive compilation of risk factors would be invaluable in identifying and preventing new causes of NDDs, and would also provide great opportunities for further systematic or personalized analysis of the genesis and progression of NDDs. In epidemiology, risk factors are variables associated with an increased risk of disease or infection, whereas protective factors are associated with a decreased risk of disease or infection [20]. Aging is a main risk factor for most NDDs [3,21] and basically there is no effective treatments for age-related NDDs. The pathologic changes that accompanied with the onset of NDDs are irreversible. NDDs are incurable and debilitating diseases, the prevention stage is critical in NDDs, and the collection of risk factors for NDDs may build confidence in both disease prevention and research.

We searched existing NDDs related databases and found a few databases sporadically included the risk factors information, such as the integrated neurodegenerative disease (INDD) database [22], which contains demographic data, as well as details of clinical assessments, neuropathological testing, imaging, pathology, biological fluids, genetics and clinical trials. The Alzheimer's Research Forum website (http://www.alzforum.org) is an independent research project and designed to develop online community resources to manage scientific knowledge, information and data about AD. The AlzRisk database (http://www.alzrisk.org/) contains epidemiological risk factors for AD [23]. The etiology of PD is still unknown and likely multifactorial. Several PD-related databases have been established in order to study the underlying causes. In PDGene [24] (http://www.pdgene.org), data from all genetic association studies published in the field of non-Mendelian PD, including GWAS, have been collected and *meta*-analyzed. PDbase [25] (http://bioportal.kobic.re.kr/PDbase/) is the database of PD-related gene and genetic variations, assembled using data from substantia nigra in PD and normal tissues. ParkDB [26] (http://www2.cancer.ucl.ac.uk/Parkinson_Db2/), which is dedicated to gene expression in PD, contains a complete set of re-analyzed, curated and annotated microarray datasets. However, there is no specific databases or knowledge base for risk factors of pan-NDDs for public use.

A database that integrates the knowledge of NDD risk factors will benefit the systematic understanding of the heterogeneity of the complex NDDs [27,28]. We hereby established a knowledge base, NDDRF, which covers different types of NDDs risk factors, such as, epidemiological factors, genetic factors, biochemical factors, combination factors *etc*., developing a powerful resource for comparing of risk factor profiles in pan-NDDs level and for the personalized prevention of NDDs. We collected data published between 1975 and 2020 and found 998 risk factors screened from 24,338 PubMed records. The risk factors in our knowledge base can be classified, displayed, searched in various ways and then analyzed. NDDRF contains substantial useful information and can be easily used to explore NDDs associated risk factors and other information, according to the needs of the investigators.

Materials and methods

Before data collection, the medical, life science and information experts as well as the potential users of the data were consulted and discussed for the scope of data and the classification of risk factors. After development of the preliminary collection plan, we modified and improved the process during the collection of data, based on the article content. In the end, we built a comprehensive data collection pipeline and developed the classification criteria.

Data inclusion criteria

Conclusion contains a clear positive statement, for example:

Risk factor: 'is a risk factor for PD', 'increases PD risk', 'could be a risk factor for PD', 'association with PD', 'causes PD', 'accelerates PD progression', *etc.*

Protective factor: 'is a protective factor for PD', 'has a positive effect on PD', *etc.*

For the comprehensiveness of the data, we included the risk factors which are described as "could be a risk factor", and also labeled them clearly as "could be a risk factor" in the database.

Data exclusion criteria

No conclusive statement, for example:

'is not a risk factor for PD', 'is not a genetic factor for PD', 'is not associated with PD' and 'is not a protective factor'.

Reviews, comments, abstracts without full text available from the original publications and book chapters were excluded.

Data sources

We searched the keywords 'neurodegenerative diseases' in the MeSH database (https://www.ncbi.nlm.nih.gov/mesh) and retrieved a total of 78 NDDs. At the same time, using the specific keywords such as, (Disease name[tiab] OR Abbreviation[tiab] OR Alias[tiab]) AND (risk factor*[tiab] OR genetic factor*[tiab] or risk marker*[tiab]), retract the literature. Meanwhile, the 78 NDDs were searched independently in PubMed and the extracted data were then mined manually.

From data deposited between 1975 to December 2020, 24,338 original articles were retrieved. In order to obtain accurate and original data, only publications with complete information were collected, the reviews, comments and book chapters (n = 8011) were not included, 16,327 articles left. For further reference, articles described animal experiments were included and marked specifically. We excluded the abstraction of these without full texts available from the original publications. Since we used abbreviations for disease names in the keywords for searching, for instance, Parkinson's disease was abbreviated to 'PD', but also referred to something else, such as peritoneal dialysis, pancreaticoduodenectomy, packing density, programmed death 1 and programmed death-ligand 1, *etc*. The articles without clear conclusive statement were further excluded. Finally 998 risk factors, 2293 records and 1071 articles relevant to 14 types of NDDs were obtained. A flowchart about details of the literature collection process is provided in Fig. 1. The same risk factor might be described differently in separate articles, so we unified the names of the risk factors.

Classification criteria

Because of the obvious difference in classification of risk factors for different diseases, we reviewed the literature and found that there is currently no standard classification for neurodegenerative risk factors. To carry out a reasonable classification, we developed the criteria for the classification of neurodegenerative risk factors by discussion with clinicians and bioinformatics experts, *etc*.[29]. The risk factors for NDDs were divided into five categories, *i.e.,* epidemiological factors, genetic factors, biochemical factors, combination factors and non-risk factors. Epidemiological factors mainly include age, gender, ethnicity, disease and some environmental factors. Genetic factors mainly include genes, gene variants and family history. Biochemical factors are various biochemical indicators. Combination factors involve the combination of factors at the

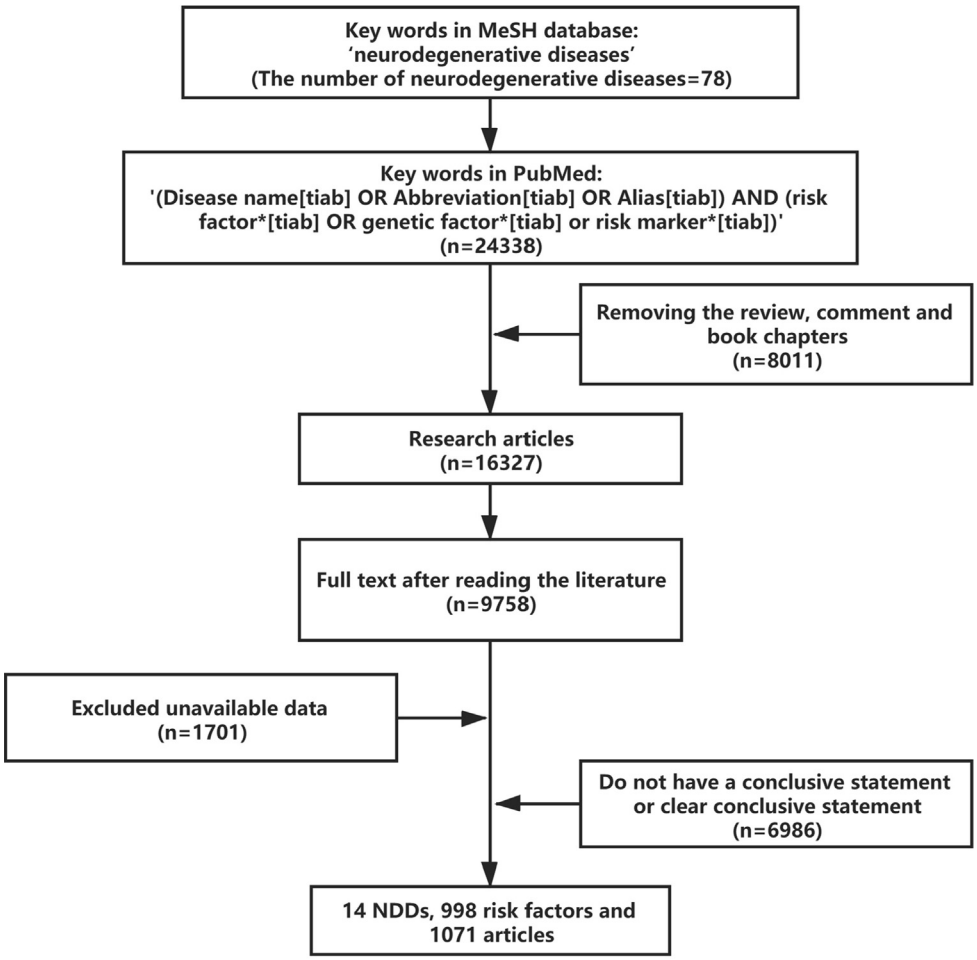


Fig. 1. Flowchart for collection of risk factors for neurodegenerative diseases.

Table 1
NDDs risk factor categories in NDDRF knowledge base.

Categories	Explanation
Epidemiological factors	Age, gender, ethnicity, disease, medication, diet, exposure to the environment, smoking, sleep, comorbidity *etc.*
Genetic factors	genes, mutations, family history, etc.
Biochemical factors	Biochemical indicators
Combination factors	Combination of epidemiological, genetic or biochemical factors
Non-risk factors	It is associated but not a risk for NDDs under certain preconditions

Note: Genetic mutations were divided into three categories: single nucleotide polymorphisms, copy number variations, and other polymorphisms.

Table 2
Risk factor information.

Attributes	Description	Examples
RF_ID	Risk factor ID	
Disease name	Disease name for particular risk factor	Parkinson's disease
RF_Category	Types of risk factors	Genetic factors
Polymorphisms category	Types of polymorphisms	Single nucleotide polymorphism
Gene	Gene name	PARK16
Variable sites	Variable sites	rs947211
Species	Corresponding species	human
Quantity	Quantification of specific value of risk factor	$\leq$ 12 years
Object	Object of risk factor	Late-onset PD
Group	Population (or race) of risk factor	Han Chinese population
Condition	Prerequisites for becoming a risk factor	Taking creatine, together with high levels of caffeine intake
Gender	Gender-related risk factor	Male or female
Age	Age-related risk factor	$\geq$ 60
Association	Risk factor or protective factor	PF or RF
Key Sentence	Concluding key sentences in the article	

Note: PF (Protective factor); RF (Risk factor)

Table 3
Statistical description of the risk factor.

Attributes	Description	Examples
Type	Experiment type	Meta-analysis
Total	Total number of people	358
Race (Region)	Ethnicity or region of experimental population	Japanese population
Case	Case information (number, gender, age)	192 (male: 114; female: 78; 66.36 ± 9.74)
Control	Control information (number, gender, age)	193 (male: 112; female: 81; 66.94 ± 9.12)
OR, aOR, HR, aHR, RR, IRR	OR (95% CI)	1.26 (1.04–1.54)
P-value	P-value	0.018
Family history	Family history	Non-hereditary idiopathic PD
Comorbidity	Comorbidity	Gaucher disease
Life Style	Life style	Creatine (5 g twice daily) or placebo for 5 years; caffeine intake during the previous week
Environment	Environment	Pesticide exposure
Country	Experimental country	France
Year	Experimental year	2000–2010

same or the different classification described above. Non-risk factors are those factors associated but not risk for NDDs. The classification are shown in Table 1.

Data model

Using the collected data as above, we created three tables for the knowledge base: risk factor information (Table 2), statistical description of the risk factors and sample (Table 3), and reference information (Table 4).

Implementation

The functionality, interface and content of the knowledge base were designed based on the application scenarios. The WAMP (Windows + Apache + MySQL + PHP) development environment was applied to the establishment of our knowledge base, with MySQL as the database server, Apache as the web server software, Hypertext Preprocessor (PHP) as the scripting language and Windows as the operating system. The WAMP development environment we used is particularly suitable for the development of small- and medium-sized databases. The entity relationships of the NDDRF knowledge base are shown as unified modeling language class diagrams in Fig. 2.

Table 4
Information for references.

Attributes	Description	Examples
PMID	PMID	21,059,511
Year	Year of article	2010
Text type	Text type	experiment
Title	Title of article	

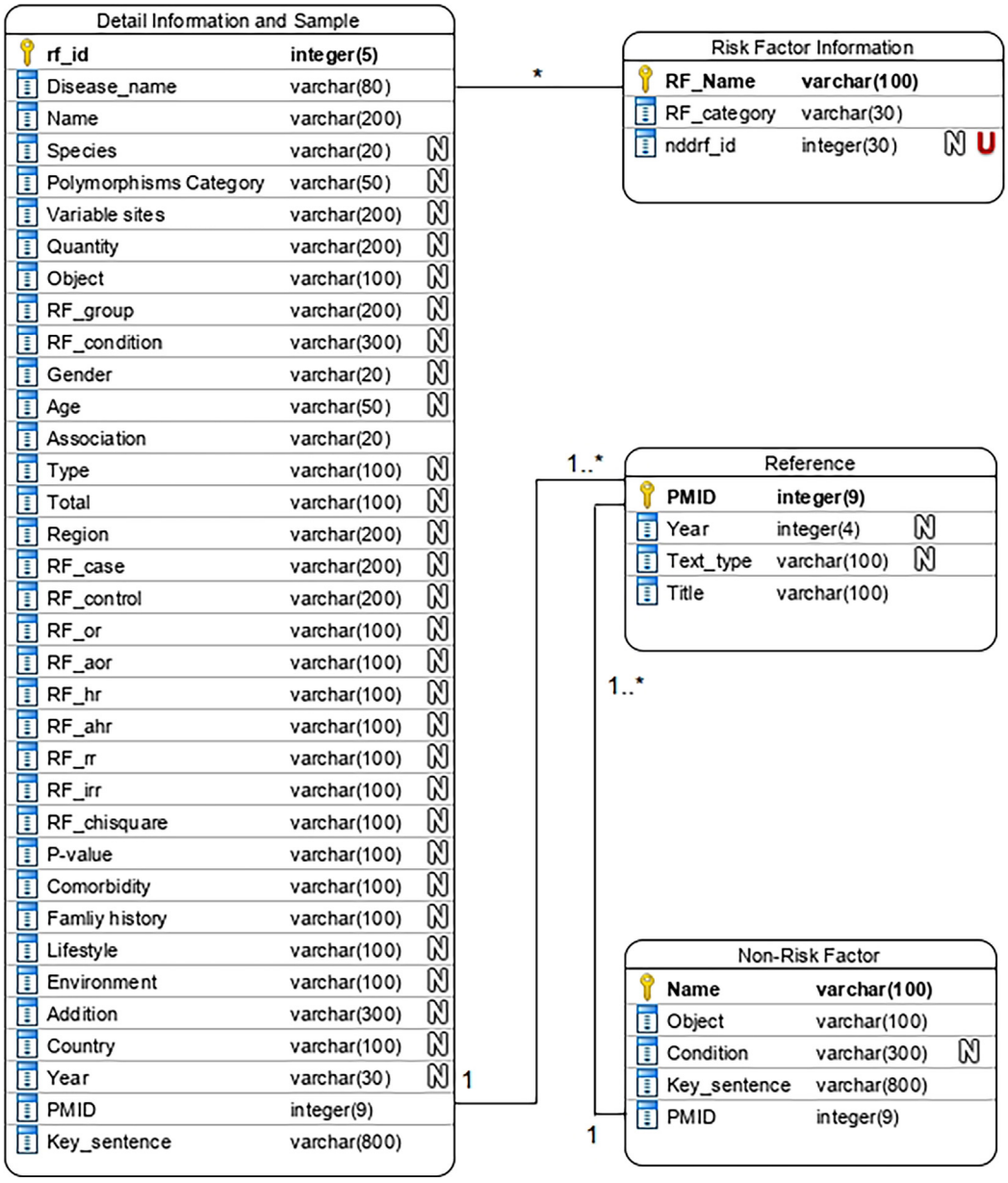


Fig. 2. Unified modeling language class diagrams of NDDRF. Note: N (allowed to be Null); U (Unique key).

Table 5
Number of records for different NDDs.

Neurodegenerative Diseases	Epidemiological Factor	Genetic Factor	Biochemical Factor	Combination Factor
Alzheimer's disease	349	459	130	67
Amyotrophic lateral sclerosis	340	98	25	12
Creutzfeldt-Jakob disease	48	20	0	0
Chronic traumatic encephalopathy	3	0	0	0
Frontotemporal dementia	21	19	2	2
Frontotemporal lobar degeneration	8	99	8	0
Hereditary motor and sensory neuropathy	0	2	0	0
Lewy body dementia	12	21	1	0
Motor neuron disease	15	1	4	0
Multiple system atrophy	23	22	0	0
Parkinson's disease	104	125	17	2
Progressive supranuclear palsy	15	11	1	0
Tourette syndrome	32	7	0	0
Tuberous sclerosis complex	16	0	0	0

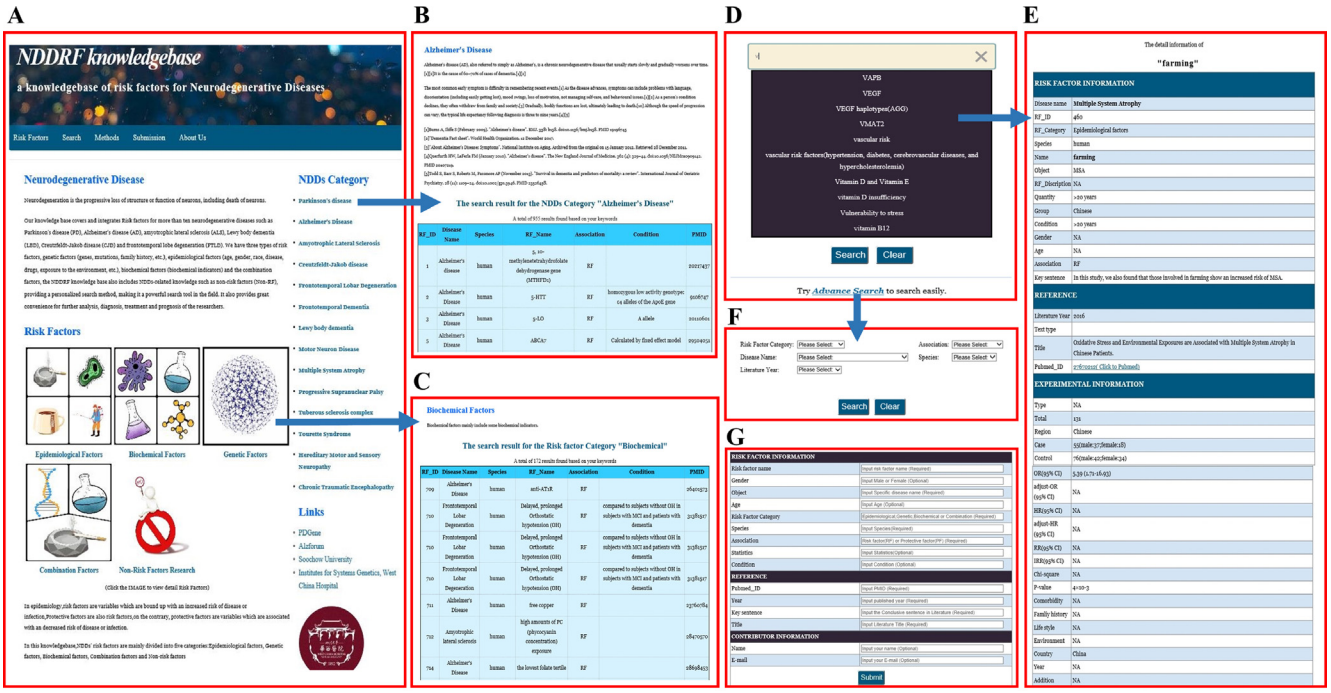


Fig. 3. Web interface of NDDRF. (A) Risk factor page; (B) Categories of NDD; (C) Categories of risk factors; (D) Search page; (E) Details of risk factor; (F) Advanced search; (G) Submission page.

Results

Data statistics

The NDDRF knowledge base comprises 14 NDDs: PD (145 risk factors, total of 248 records), AD (492 risk factors, total of 1005 records), ALS (206 risk factors, total of 417 records), Creutzfeldt-Jakob disease (27 risk factors, total of 68 records), chronic traumatic encephalopathy (three risk factors, total of three records), frontotemporal dementia (30 risk factors, total of 44 records), frontotemporal lobar degeneration (44 risk factors, total of 115 records), hereditary motor and sensory neuropathy (one risk factor, total of two records), Lewy body dementia (21 risk factors, total of 34 records), motor neuron disease (14 risk factors, total of 20 records), multiple system atrophy (18 risk factors, total of 45 records), progressive supranuclear palsy (23 risk factors, total of 27 records), Tourette's syndrome (33 risk factors, total of 39 records) and tuberous sclerosis complex (16 risk factors, total of 16 records).

NDDRF contains 946 individual risk factors and 52 combination risk factors. The individual risk factors were divided into 3 categories, based on our classification criteria: 469 epidemiological factors, total of 986 records; 324 genetic factors, total of 885 records; and 153 biochemical factors, total of 188 records. The knowledge base also contains 129 non-risk factors, with a total of 147 records. Classified by association, NDDRF contains 179 protective factors with a total of 303 records and 880 risk factors with a total of 1844 data. The details are listed in Table 5.

Database schema

The NDDRF knowledge base has 6 pages: a homepage, which gives a visual representation of the main contents of the knowledge base; a page for risk factors; a search page, which allows basic and advanced searches; a methods page, which includes an introduction to NDDRF and directions for use; a submission page, which allows users to submit new risk factors; and an 'About Us' page, which gives details of our team and the related researches.

Web interface and function

Risk factors page

The risk factors page (Fig. 3A) contains an introduction to NDDs and risk factors. The right-hand side of the home page displays the 14 NDDs in the knowledge base, as well as links of related databases and websites for users to browse. When users click on the disease name, a page will be presented to introduce the selected disease and related information about risk factor (Fig. 3B). At the lower half of the risk factors page, the 5 categories of risk factors for NDDs, *i.e.,* epidemiological factors, genetic factors, biochemical factors, combination factors and non-risk factors, will be shown as pictures. Users can click the type of risk factor that they want to browse and check the information (Fig. 3C).

Search page

The search page (Fig. 3D) allows both basic and advanced searches. The basic search facility mainly allows 'fuzzy' queries, which is suitable for direct queries of risk factor names. When users enter the keywords of a risk factor in the search box, a list of the top ten related information about suggestions for the keywords will be displayed for selection below the search box. After the keyword selected and the search button clicked, the page will display the search results and the basic information (ID, disease name, species, name, association, condition and PMID) of the risk factor. Users can click on the item they want to check (Fig. 3E). The risk factor information, reference information and experimental information, will also be presented in tables. If users want to know more about this risk factor, they can click on the PMID in the reference, and a hyperlink is access to the original article.

The advanced search facility allows combined searches. On this page (Fig. 3F), we have provided five drop-down lists for different types of query: risk factor category drop-down list (includes epidemiological factors, genetic factors, biochemical factors and combination factors), association drop-down list (includes risk factors and protective factors), disease name drop-down list (includes the 14 types of NDDs recorded in the NDDRF), species drop-down list (human and animal) and literature year (from 1984 to 2020). Therefore, users can search terms individually or in combination, in line with their personalized applications.

Methods page

On the methods page, we explain the purpose of the NDDRF knowledge base, the items in the knowledge base and the usage of the knowledge base. By this module, users can quickly understand the detail of our knowledge base and apply the knowledge base to their research.

Submission page

On the submission page (Fig. 3G), we allow users to submit new risk factors to the table. The submission form contains the following fields for the risk factor and reference: risk factor name, gender, object, age, risk factor category, species, association, statistics, condition, PMID, year, key sentence and title (the fields are explained and we indicate whether they are 'optional' or 'required'). When the user completed the form and clicked the submit button, data will be uploaded to our database immediately. The new risk factor information will be added to our NDDRF knowledge base after our manual checking and review. We also require contributors to provide their names and contact information on the submission form so that we can notify them after their data were updated into our knowledge base.

About us page

This page provides an introduction to our team and the related researches on NDDs.

Descriptive statistics

We carried out a statistical summary of the data in the NDDRF. Taking AD and PD as examples, the numbers of articles on risk factors has increased year by year (Fig. 4A), it indicates that the prevention of NDDs is seen as particularly important. In the future, more studies will be incorporated and updated into the NDDRF. Among the 5 types of risk factors, epidemiological factors are the most, followed by genetic factors (Fig. 4B). In Fig. 4C, we analyzed the proportion of protective factors and risk factors for 14 NDDs, in which protective factors accounted about 17% (Fig. 4C). Among the 14 NDDs included in the knowledgebase, AD is the most studied, followed by ALS and PD. About 1,000 researches were collected for risk factors of AD (Fig. 4D). Furthermore, the three diseases are also with the highest morbidity in NDDs. Our knowledge base includes studies from 51 countries (Fig. 4E), and most studies were carried out in the United States, followed by China, Italy, and the United Kingdom. In order to more fully understand the correlation between NDDs and risk factors, we visualized the data and made a weighted network diagram of the correlation between NDDs and risk factors (Fig. 4F). The nodes of apolipoprotein E (15), education (425), microtubule-associated protein tau (174), smoking (652)

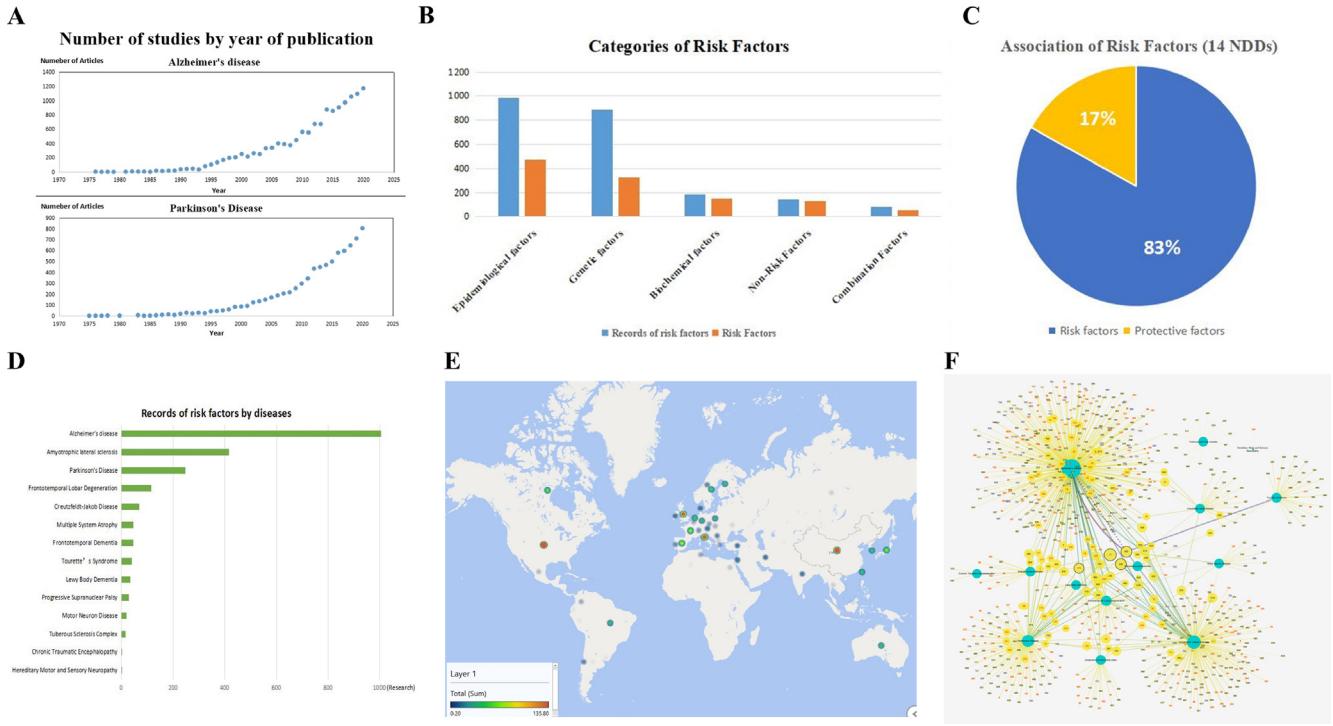


Fig. 4. Descriptive statistics for NDDRF. (A) Chart shows the number of studies by year of publication; (B) Chart shows categories of risk factors; (C) Chart shows the association of risk factors for all 14 NDDs; (D) Chart shows the records of risk factors by disease in NDDRF; (E) Heat map shows country distribution; (F) Weighted network diagram shows correlation between neurodegenerative diseases and risk factors. The nodes represent NDDs and risk factors, and number for nodes indicates the risk factor ID in the NDDRF. The size of the node for each NDD indicates the number of associated risk factors and the size of node for each risk factor represents the number of associated NDDs. The thickness and depth of the edges represent connectivity. The edge of the solid line represents risk factors and the dotted line represents protective factors. The RF nodes are colored black, red and blue to represent epidemiological, genetic and biochemical factors.

and obesity (3 1 5) are correlated with a variety of NDDs. These risk factors have been studied extensively and are common risk factors for many NDDs, which shows that these top risk factors are very important.

Discussion

Although more and more studies on the pathogenesis of NDDs and the risk factors are intensively studied and discovered, recently the interaction between risk factors and pathogenic mechanisms requires further study. Because the pathological changes associated with the onset of most NDDs are irreversible, patients often have cognitive dysfunction in the middle and the late stages of the disease. Since there are no effective treatments to delay the progression of NDDs, the latent stage of the disease is the most valuable and effective stage to begin neuroprotective therapy and control. With the paradigm shifting from clinical management to healthcare, prevention and risk assessment of NDDs are becoming very necessary [30,31]. We investigated several public databases of NDDs, including databases on gene expression and genetic variations associated with PD, such as PDGene, PDbase and ParkDB, and AlzRisk, which is a comprehensive database covering 15 epidemiological risk factors for AD. There is currently neither database of PD risk factors, nor a comprehensive database covering other types of risk factors.

With the progression of TCGA (The Cancer Genome Atlas) project, comparison of molecular alteration profiles of similar diseases is becoming a new research paradigm [32–36]. NDDs are a series of disorders progressed with the structural and functional degeneration of the central or peripheral nervous system. The mechanisms for the genesis and progression of these disease could be similar [37–40]. A knowledge base of risk factors for NDDs would play an important role to the pan-NDDs study. We have previously studied the similarity of mutation profiles for NDDs [13] and biomarker profiles for cardiovascular diseases [41], some interesting phenomenon has been observed with the pan-disease analysis. As seen from Fig. 4 (F), the risk factors for the different NDDs shared several same key risk factors, which can only be discovery by the pan-NDDs analysis. The further study is expected to investigate the common risk factors for understanding their molecular mechanisms.

NDDRF is the first knowledge base containing risk factors for NDDs and provides users with accurate and comprehensive risk factor information. The data were collected and filtered by our professional and experienced researchers. On the basis of the original keyword search, we also conducted a secondary search for risk factors that were rarely reported in the results. Our database thus contains comprehensive information about risk factors and includes data from a wide range of original documents. NDDRF includes 324 genetic factors, 469 epidemiological factors (including environmental factors), 153 biochemical factors and 52 combination factors from the classification of risk factors, it also includes 179 protective factors and 880 risk factors. We covered experimental information, risk factor-related information and literature information for 14 NDDs.

We compared the risk factors of AD in NDDRF with those in the AlzRisk database. Interestingly, we found that head injury and obesity are the shared top risk factors. However, AlzRisk only contains epidemiological risk factors and was last updated in 2018, many of the top risk factors in our knowledgebase are therefore not included in AlzRisk. NDDRF has several advantages compared with other NDD databases (Table 6). Firstly, the types of risk factors included in NDDRF are more comprehensive, some non-risk factors and risk models are also incorporated, and risk factors and protective factors are distinguished. Secondly, compared with many machine mining databases, manually collected data is more accurate and comprehensive. Thirdly, the search capabilities that we provide include fuzzy search and list search, which makes NDDRF more convenient to use. Since knowledgebase is important for the knowledge graph extraction and the future explainable artificial intelligence (AI) modeling [42]. NDDRF could be an important resource to build the trustable and explainable AI for precision diagnosis and prevention of NDDs.

So far, we only collected risk factors from PubMed and not from other databases such as EMBASE, *etc*., in other words, some data conforming to our inclusion standards were not included in our knowledge base.

In the next version of the database updating, we will try to improve the diversity and quality of the data collected in the database. With the fast developing of modern high throughput technologies and the smart sensors, many types of new risk factors related to NDDs will be explored and reported. We are planning to incorporate all the novel risk factors to our knowledgebase. In addition, if one risk factor is reported by only one or few publications, the reliability of the risk factor may be lower than the one with many publications and supports. A quality score will be developed to access the reliability of the risk factors, considering the journal's impact, the citation of the work and the sample size used in the research, *etc*.

Furthermore, the NDD level ontology could be built to standardize the NDD terminologies and their relationships [43,44]. The NDD ontology can integrate the International Classification of Diseases (ICD) codes, the MeSH terms, the knowledge in our database and other useful NDD information to build a set of concepts, relationships for the sharing of NDD data and for the artificial intelligent modellings for NDD classification and personalized prevention.

Table 6
Comparison of NDD databases.

	PDGene	PDbase	AlzRisk	NDDRF
Purpose of the database	A database for collecting and *meta*-analyzing data from all genetic association studies published in the field of non-Mendelian PD	A database of PD-related gene and genetic variations	A database for collecting epidemiologic reports that evaluate environmental risk factors for AD in well-defined study cohorts	A knowledgebase for collecting different categories of risk factors related to 14 NDDs from literature
Data resource	Scientific literature and other scientific resources	Scientific literature and other scientific resources	Scientific literature	Scientific literature
Data collection	Manually curated	Manually curated	Manually curated	Manually curated
Data type	GWAS	Data from substantia nigra in PD and normal tissues	Epidemiological factors	Epidemiological factors, Genetic factors, Biochemical factors and Combination factors

Conclusion

We constructed a knowledge base of risk factors for NDDs, NDDRF, for the future systematic understanding, diagnosis and prevention of NDDs. With NDDRF as a knowledge resource, we will continuously update our knowledge base, incorporate visual statistical functions and build personalized models to predict the risk of NDDs.

Most of the present AI systems are nonlinear, complex and black-box models, knowledgebase is human intelligence (HI) based, combining AI with HI is very important in the medicine and healthcare field, NDDRF knowledgebase provides the personalized knowledge, which could be utilized as structured human knowledge for the construction of knowledge graphs and be implemented to the AI models and promotes the development of explainable AI for the NDD diagnosis and prevention.

Notes:

Database URL: http://sysbio.org.cn/NDDRF/.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2016YFC1306605), the National Natural Science Foundation of China (32070671) as well as the regional innovation cooperation between Sichuan and Guangxi Provinces (2020YFQ0019).

References

[1] Appel SH, Smith RG, Le WD. Immune-mediated cell death in neurodegenerative disease. Adv Neurol. 1996;69:153–9.
[2] Hardy J. Pathways to primary neurodegenerative disease. Ann N Y Acad Sci. 2000;924:29–34.
[3] Kritsilis M, S VR, Koutsoudaki PN, Evangelou K, Gorgoulis VG, Papadopoulos D. Ageing, Cellular Senescence and Neurodegenerative Disease. Int J Mol Sci. 2018;19(10).
[4] Programme EJ. Neurodegenerative Disease Research [Internet]. "What is Neurodegenerative Disease?" JPND Research. 2019. Available from: https://www.neurodegenerationresearch.eu/about/what/.
[5] Gammon K. Neurodegenerative disease: brain windfall. Nature 2014;515(7526):299–300.
[6] Hawkes CH, Del Tredici K, Braak H. A timeline for Parkinson's disease. Parkinsonism Relat Disord. 2010;16(2):79–84.
[7] Gitler AD, Dhillon P, Shorter J. Neurodegenerative disease: models, mechanisms, and a new hope. 2017;10(5):499–502.
[8] Christina, P. World Alzheimer's report 2018 Alzheimer's disease internations: world alzheimer report 2018.: 2018.
[9] Rubinsztein DC. The roles of intracellular protein-degradation pathways in neurodegeneration. Nature. 2006;443(7113):780-786.
[10] Bredesen DE, Rao RV, Mehlen P. Cell death in the nervous system. Nature 2006;443(7113):796–802.
[11] Labadorf A, Choi SH, Myers RH. Evidence for a Pan-Neurodegenerative Disease Response in Huntington's and Parkinson's Disease Expression Profiles. Front Mol Neurosci. 2017;10:430.
[12] Noori A, Mezlini AM, Hyman BT, Serrano-Pozo A, Das S. Systematic review and meta-analysis of human transcriptomics reveals neuroinflammation, deficient energy metabolism, and proteostasis failure across neurodegeneration. Neurobiol Dis. 2020;149:105225.
[13] Yang Y, Xu C, Liu X, Xu C, Zhang Y, Shen L, et al. NDDVD: an integrated and manually curated Neurodegenerative Diseases Variation Database. Database (Oxford) 2018;2018.
[14] Yang Z, Li T, Cui Y, Li S, Cheng C, Shen B, et al. Elevated Plasma microRNA-105-5p Level in Patients With Idiopathic Parkinson's Disease: A Potential Disease Biomarker. Front Neurosci. 2019;13:218.
[15] Yang Z, Li T, Li S, Wei M, Qi H, Shen B, et al. Altered Expression Levels of MicroRNA-132 and Nurr1 in Peripheral Blood of Parkinson's Disease: Potential Disease Biomarkers. ACS Chem Neurosci. 2019;10(5):2243–9.
[16] Li T, Yang Z, Li S, Cheng C, Shen B, Le W. Alterations of NURR1 and Cytokines in the Peripheral Blood Mononuclear Cells: Combined Biomarkers for Parkinson's Disease. Front Aging Neurosci. 2018;10:392.
[17] Singla RK, Agarwal T, He X, Shen B. Herbal Resources to Combat a Progressive & Degenerative Nervous System Disorder - Parkinson's Disease. Curr Drug Targets 2020.
[18] Shen B, Lin Y, Bi C, Zhou S, Bai Z, Zheng G, et al. Translational Informatics for Parkinson's Disease: from Big Biomedical Data to Small Actionable Alterations. Genomics Proteomics Bioinformatics. 2019;17(4):415–29.
[19] Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test?. BMJ 1999;319(7224):1562–5.
[20] H PR, F. TM. Disorders of childhood: Development and psychopathology. Nelson Education2013.
[21] Hou Y, Dan X, Babbar M, Wei Y, Hasselbalch SG. Ageing as a risk factor for neurodegenerative disease. 2019;15(10):565–81.
[22] Xie SX, Baek Y, Grossman M, Arnold SE, Karlawish J, Siderowf A, et al. Building an integrated neurodegenerative disease database at an academic health center. Alzheimers Dement. 2011;7(4):e84–93.
[23] Kinoshita J, Clark T. Alzforum. Methods Mol Biol. 2007;401:365–81.
[24] Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide BM, et al. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. PLoS Genet. 2012;8(3):e1002548.
[25] Yang JO, Kim WY, Jeong SY, Oh JH, Jho S, Bhak J, et al. PDbase: a database of Parkinson's disease-related genes and genetic variation using substantia nigra ESTs. BMC Genomics. 2009;10 Suppl 3(Suppl 3):S32.
[26] Taccioli C, Tegner J, Maselli V, Gomez-Cabrero D, Altobelli G, Emmett W, et al. ParkDB: a Parkinson's disease gene expression database. Database (Oxford) 2011;2011:bar007..
[27] Zhang X, Sun XF, Cao Y, Ye B, Peng Q, Liu X, et al. CBD: a biomarker database for colorectal cancer. Database (Oxford) 2018;2018.
[28] Zhan C, Shi M, Wu R, He H, Liu X, Shen B. MIRKB: a myocardial infarction risk knowledge base. Database (Oxford) 2019;2019.
[29] J J, S C, L C, W W, P X, B L, et al. Neurology. People's Medical Publishing House (PMPH); 2018.
[30] Bai J, Shen L, Sun H, Shen B. Physiological Informatics: Collection and Analyses of Data from Wearable Sensors and Smartphone for Healthcare. Adv Exp Med Biol. 2017;1028:17–37.
[31] Shen L, Ye B, Sun H, Lin Y, van Wietmarschen H, Shen B. Systems Health: A Transition from Disease Management Toward Health Promotion. Adv Exp Med Biol. 2017;1028:149–64.
[32] Fehlmann T, Lehallier B, Schaum N, Hahn O, Kahraman M, Li Y, et al. Common diseases alter the physiological age-related blood microRNA profile. Nat Commun. 2020;11(1):5958.
[33] Li S, Li J, Chen C, Zhang R, Wang K. Pan-cancer analysis of long non-coding RNA NEAT1 in various cancers. Genes Dis. 2018;5(1):27–35.
[34] Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 2018;173(2):283–5.
[35] Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science 2018;362(6411).
[36] Yau NK, Fong AY, Leung HF, Verhoeft KR, Lim QY, Lam WY, et al. A Pan-Cancer Review of ALK Mutations: Implications for Carcinogenesis and Therapy. Curr Cancer Drug Targets. 2015;15(4):327–36.
[37] Hussain M, Kumar P, Khan S, Gordon DK, Khan S. Similarities Between Depression and Neurodegenerative Diseases: Pathophysiology, Challenges in Diagnosis and Treatment Options. Cureus. 2020;12(11):e11613.
[38] Melki R. Role of Different Alpha-Synuclein Strains in Synucleinopathies, Similarities with other Neurodegenerative Diseases. J Parkinsons Dis. 2015;5(2):217–27.
[39] Letiembre M, Liu Y, Walter S, Hao W, Pfander T, Wrede A, et al. Screening of innate immune receptors in neurodegenerative diseases: a similar pattern. Neurobiol Aging. 2009;30(5):759–68.
[40] Ridley RM, Baker HF, Crow TJ. Transmissible and non-transmissible neurodegenerative disease: similarities in age of onset and genetics in relation to aetiology. Psychol Med. 1986;16(1):199–207.
[41] Wu R, Lin Y, Liu X, Zhan C, He H, Shi M, et al. Phenotype-genotype network construction and characterization: a case study of cardiovascular diseases and associated non-coding RNAs. Database (Oxford) 2020;2020.
[42] Ammar N, Shaban-Nejad A. Explainable Artificial Intelligence Recommendation System by Leveraging the Semantics of Adverse Childhood Experiences: Proof-of-Concept Prototype Development. JMIR Med Inform. 2020;8(11):e18752.
[43] Shen L, Bai J, Wang J, Shen B. The fourth scientific discovery paradigm for precision medicine and healthcare: challenges ahead. Precision. Clinical Medicine. 2021.
[44] Chen Y, Yu C, Liu X, Xi T, Xu G, Sun Y, et al. PCLiON: An Ontology for Data Standardization and Sharing of Prostate Cancer Associated Lifestyles. Int J Med Inform. 2021;145:104332.