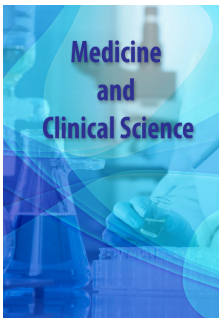
Medicine & Clinical Science

Original Article



Elucidation of the mechanism of cognitive decline in mild cognitive impairment using causal discovery

Kenji Matsuura

Faculty of Pharmacy, Osaka-Ohtani University, Tondabayashi , Japan

Correspondence

Kenji Matsuura
Faculty of Pharmacy, Osaka-Ohtani University, Tondabayashi 584-8540, Japan
Tel: 81-042-635-2551
Fax: 81-042-635-2661
Email: matuuken@osaka-ohtani.ac.jp

Abstract

Mild cognitive impairment (MCI) is the stage between the expected cognitive decline of normal aging and the more serious decline of dementia. Since MCI may increase risk of later developing dementia caused by Alzheimer's disease (AD) or other neurological conditions, it is important to detect and treat MCI early in order to prevent dementia. Causal discovery, which can visualize causal relationships (cause and effect) among data, has recently been attracting attention. One of its algorithms, the Linear Non-Gaussian Acyclic Model (LiNGAM), can extract causal relationships among variables from statistical data only, using probability distributions of variables that are generally non-Gaussian. In this study, we used LiNGAM to analyze gene expression data in the hippocampus of healthy subjects and of MCI patients, and also the Mini-Mental State Examination (MMSE) scores, which are used to assess cognitive decline. We found that cell adhesion molecule 4 (CADM4) gene regulates cognitive processes that are measured in terms of MMSE scores. Our results revealed a causal relationship between gene expression changes and MMSE scores, and allowed us to identify the genes responsible for cognitive decline..

- Received Date: 08 Jun 2022
- Accepted Date: 01 Jul 2022
- Publication Date: 05 Jul 2022

Introduction

Alzheimer's disease (AD) accounts for 60-70% of the estimated 50 million people worldwide suffering from dementia [1]. AD is a neurodegenerative disease of the dementia type that is associated with cognitive dysfunction, functional changes, and changes in brain organization such as reduced hippocampal volume. The cause of AD is thought to be the accumulation of amyloid-β (Aβ) and hyperphosphorylated tau protein in the brain. Mild cognitive impairment (MCI) is a cognitive condition intermediate between normal aging and early dementia, in which short-term memory loss and other symptoms due to cognitive decline are observed but do not reach the level of dementia and do not interfere with daily life. MCI is found in approximately 16% of the elderly population aged 65 years or older and is considered a "pre-dementia" condition [2]. Although 5-15% of MCI patients develop dementia each year, this does not necessarily mean that all MCI patients will develop dementia [3]. Also, it has been reported that 14-44% of patients with MCI can be restored to health with exercise and cognitive training [3]. Since recovery from dementia to MCI is not possible, it is important to detect and treat MCI at the MCI stage in order to prevent dementia.

Machine learning models such as multiple regression models, support vector machines, and random forests are used to analyze big data, but these correlation analyses cannot reveal causal relationships among variables. Therefore, in recent years, causal structure search, which can visualize causal relationships (cause and effect) among data, such as which variable generated which variable's data, has been attracting attention. One such algorithm, the Linear non-Gaussian Acyclic Model (LiNGAM), can extract causal relationships among variables from statistical data alone, using probability distributions of variables that are generally non-Gaussian [4]. Direct LiNGAM, one of the LiNGAM algorithms, extracts causal structure from multivariate data using single regression and independence evaluation [5]. It then visualizes the causal structure by repeatedly computing single regression and independence evaluation among variables until the ordinal relationship among variables in the data is determined.

Clarification of the causes of cognitive decline in MCI will lead to elucidation of the mechanisms of AD onset and establishment of AD prevention methods. In this study, I applied hippocampal microarray data from healthy and MCI subjects to the Direct LiNGAM algorithm to identify genes responsible for cognitive decline in MCI and visualize their molecular mechanisms.

Citation: Matsuura K. Elucidation of the mechanism of cognitive decline in mild cognitive impairment using causal discovery. Med Clin Sci. 2022; 4(2):1-6.

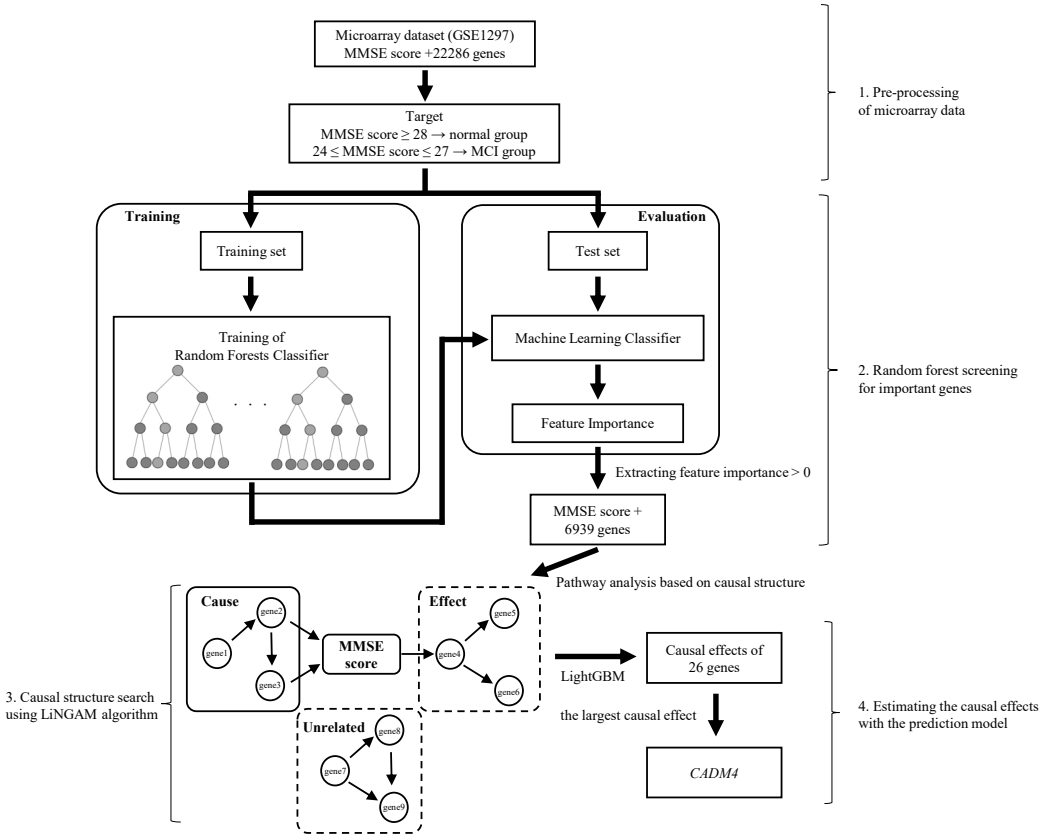


Figure 1. *Flowchart of the study, representing the major steps of data manipulation, as outlined in Materials and Methods.*

Methods

Data analysis and calculations were performed in four steps (Figure 1).

Pre-processing of microarray data

The microarray dataset GSE1297 was retrieved from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo). GSE1297 is a dataset containing gene expression data of human hippocampus from control subjects and AD subjects and their MiniMental State Examination (MMSE) scores. MMSE score is a reliable indicator of AD-related cognitive status at a point in time, defined as follows: dementia group (0 ≤ MMSE score ≤ 23), MCI group (24 ≤ MMSE score ≤ 26) and normal group (27 ≤ MMSE score ≤ 30) [6]. Based on the MMSE scores, the subject data were classified into normal, MCI, and dementia groups, and the normal group and the MCI group data were extracted from them.

Random forests screening for important genes

Random forests (RF) are popular tree-based ensemble machine learning algorithms. The data from normal subjects and MCI patients were randomly split into training (70%) and test (30%) sets. The model was trained on a training set with 10,000 trees, followed by a test set to measure prediction accuracy by the RF classifier from scikit-learn package in Python (https://scikit-learn.org/). The extraction of important genes was carried out using feature importance values built in the RF algorithm [7]. In a tree-based model, each node divides the data by the feature with the greatest improvement in Gini impurity from the parent node. The Gini impurity of a given node k is defined as follows:

$$G(k) = \sum_{i=1}^{n} p(i) \times (1 - p(i)),$$

where $G(k)$ is the Gini impurity at node k, n is the number of target labels, and $p(i)$ is the frequency of target label i at node k.

The feature importance (Gini importance) of the gene j is defined as the sum of the impurity improvements of the nodes using the gene:

$$I(j) = \sum_{i=1}^{n \in F(j)} (N_{parent}(i) \times G_{parent}(i)) - (N_{left\_child}(i) \times G_{left\_child}(i) + N_{right\_child}(i) \times G_{right\_child}(i)),$$

where $I(j)$ is the feature importance of the gene j, $F(j)$ is the set of nodes for which the gene j is to be split, $N_{parent}(i)$ is the number of samples at node i, $N_{left\_child}(i)$ is the number of samples from the left node among the child nodes of node i, $N_{right\_child}(i)$ is the number of samples from the right node among the child nodes of node i, $G_{parent}(i)$ is the Gini impurity at node i, $G_{left\_child}(i)$ is the Gini impurity of the left node among the child nodes of node i, and $G_{right\_child}(i)$ is the Gini impurity of the right node among the child nodes of node i.

The genes were ranked by feature importance values, and the 6,939 genes with feature importance values greater than 0 were considered important genes.

Causal structure search using LiNGAM algorithm

First, the following equation was used to normalize the values of the dataset to a range of 0 to 1:

$$z_i = \left(x_i - \min(\mathrm{x})\right) / \left(\max(\mathrm{x}) - \min(\mathrm{x})\right),$$

where z_i is the i^{th} normalized value in the dataset, x_i is the i^{th} value in the dataset, and min(x) and max(x) are the minimum and maximum values in the dataset, respectively.

Second, as a prior knowledge, the column of MMSE scores in the dataset was specified to be the sink variable. Prior knowledge was used to estimate causal order and prune redundant edges. Based on prior knowledge, the normalized dataset was analyzed with the Direct LiNGAM algorithm to visualize the causal structure of genes that give rise to altered MMSE scores.

Estimating the causal effects with the prediction model

First, a single regression model was created in Light Gradient Boosting Machine (LightGBM). LightGBM is a framework developed by Microsoft Corporation to handle the Gradient Boosting Decision Tree algorithm (https://github.com/Microsoft/LightGBM). Next, causal effects were estimated with a predictive model. Estimates of causal effects were calculated using the following equation:

$$\mathrm{Effect}_{\mathrm{plus}} = \mathrm{E}\left[\mathrm{MMSE\ score}|\mathrm{do}\left(gene_i = \mathrm{mean}\right)\right] - \mathrm{E}\left[\mathrm{MMSE\ score}|\mathrm{do}\left(gene_i = \mathrm{mean} + \mathrm{std}\right)\right]$$

$$\mathrm{Effect}_{\mathrm{minus}} = \mathrm{E}\left[\mathrm{MMSE\ score}|\mathrm{do}\left(gene_i = \mathrm{mean}\right)\right] - \mathrm{E}[\mathrm{MMSE\ score}\,|\,\mathrm{do}(gene_i = \mathrm{mean} - \mathrm{std})],$$

where E[MMSE score|do($gene_i$ = mean)] is the expected value of an MMSE score when the value of $gene_i$ is the mean, E[MMSE score|do($gene_i$ = mean + std)] is the expected value of the MMSE score when the value of $gene_i$ is the mean + std, and E[MMSE score|do($gene_i$ = mean – std)] is the expected value of the MMSE score when the value of $gene_i$ is the mean - std. The causal effect indicates how much the expected value of the MMSE score changes when the value of $gene_i$ is changed from the mean value to the mean + std or the mean - std. The gene with the largest Effect value is the gene with the largest causal effect.

Results

Extraction of important genes contributing to the development of MCI by the Gini impurity-based feature importance of Random Forests Classifier

The GEO microarray dataset (GEO accession GSE1297) contained gene expression data in the hippocampus of healthy subjects and AD patients and their MMSE scores. This dataset was divided into a healthy subject group (7 subjects), an MCI group (7 subjects), and a dementia group (17 subjects) based on the MMSE scores. To identify genes that may regulate MCI development in the hippocampus, the dataset GSE1297 was analyzed using RF and then feature importance values based on Gini impurity incorporated into the RF algorithm were used to extract important genes that contribute to MCI development. This analysis resulted in the identification of 6,939 genes according to a threshold of feature importance values greater than 0.

Identifying causative genes that alter MMSE scores using causal discovery

The Direct LiNGAM algorithm is a new method for estimating structural equation models. It is based on exploiting the linearity and non-Gaussianity of the data [8]. The expression levels of 6,939 identified genes were normalized and converted to a range of 0 to 1, then analyzed by the Direct LiNGAM algorithm. Twenty-seven genes responsible for differences in MMSE scores were identified (Table 1) and their causal structures were visualized (Figure 2). The arrows represent causal relations between two connected nodes.

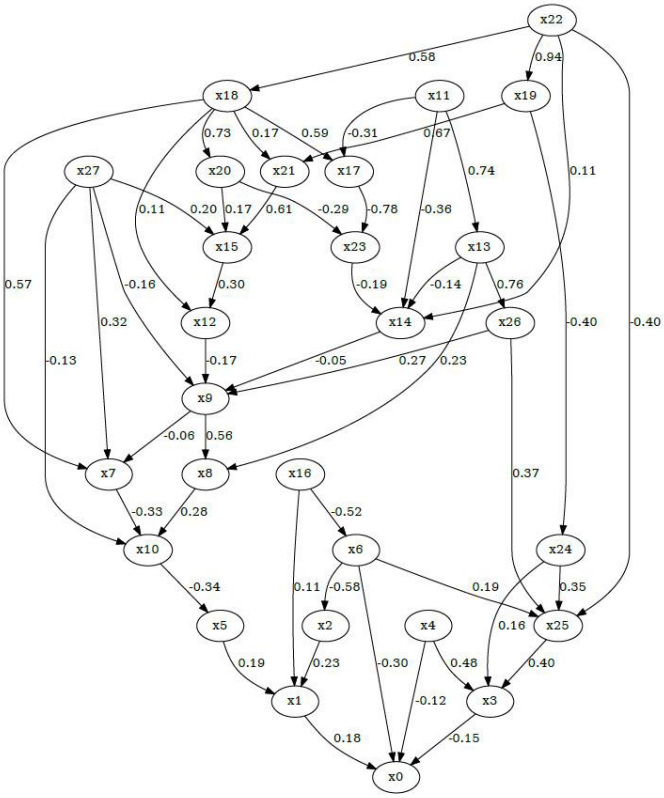


Figure 2. *The estimated network as a causal structure that changes the MMSE scores by the Direct LiNGAM algorithm. The node numbers correspond to the Node Numbers in Table 1. Arrows indicate causal relationships between two connected nodes, where the root of the arrow is the cause and the tip is the result. The numerical value of the arrow indicates the regression coefficient; a positive value indicates a positive contribution and a negative value indicates a negative contribution.*

Identification of gene with largest causal effect on MMSE scores

In order to identify the genes with the largest causal effect on MMSE scores, the causal effect of each gene on MMSE scores was estimated after creating a predictive model by LightGBM (Table 1). The gene expression changes of cell adhesion molecule 4 (*CADM4*) was found to have the largest causal effect on MMSE scores. As seen in Figure 1, the regression coefficient from *CADM4* (x2) to natriuretic peptide receptor 1 (*NPR1*) (x1) is 0.23, indicating that the increased expression of *CADM4* increases the expression of *NPR1*. Furthermore, the regression coefficient from *NPR1* (x1) to MMSE score (x0) is 0.18, indicating that increased expression of *NPR1* increases MMSE score. These results indicate that MMSE scores can be efficiently raised by increasing the expression level of *CADM4*.

Table 1. *Genes that comprise the causal structure and their causal effects on MMSE scores*

				Causal effect	
Node No.	**ID_REF**	**Gene symbol**	**Full name**	**Effect_plus**	**Effect_minus**
x0	MMSE scores	-	-	0	0
x1	32625_at	NPR1	natriuretic peptide receptor 1	0	0.28
x2	215258_at	CADM4	cell adhesion molecule 4	0	0.39
x3	213754_s_at	PAIP1	poly(A) binding protein interacting protein 1	0	0
x4	212122_at	RHOQ	ras homolog family member Q	0	0.13
x5	208668_x_at	HMGN2	high mobility group nucleosomal binding domain 2	0	0.28
x6	204198_s_at	RUNX3	runt-related transcription factor 3	0.28	0
x7	204007_at	FCGR3B	Fc fragment of IgG, low affinity IIIb, receptor (CD16b)	0	0.03
x8	202769_at	CCNG2	cyclin G2	0	0
x9	201707_at	PEX19	peroxisomal biogenesis factor 19	0	0
x10	201366_at	ANXA7	annexin A7	0	0
x11	AFFX-PheX-5_at	-	-	0.16	0
x12	36004_at	IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	0	0
x13	222351_at	PPP2R1B	protein phosphatase 2, regulatory subunit A, beta	0.16	0
x14	221712_s_at	WDR74	WD repeat domain 74	0	0
x15	221575_at	SCLY	selenocysteine lyase	0	0.01
x16	221307_at	KCNIP1	Kv channel interacting protein 1	0	0.34
x17	220935_s_at	CDK5RAP2	CDK5 regulatory subunit associated protein 2	0	0
x18	220899_at	-	-	0	0.03
x19	220862_s_at	-	-	0	0.07
x20	220759_at	EDDM3B	epididymal protein 3B	0	0
x21	220344_at	C11orf16	chromosome 11 open reading frame 16	0	0.02
x22	220318_at	EPN3	epsin 3	0	0.04
x23	220076_at	ANKH	ANKH inorganic pyrophosphate transport regulator	0	0
x24	219745_at	TMEM180	transmembrane protein 180	0.16	0
x25	219574_at	MARCHF1	membrane-associated ring finger 1	0.16	0
x26	218784_s_at	SAYSD1	SAYSVFN motif domain containing 1	0.16	0
x27	218739_at	ABHD5	abhydrolase domain containing 5	0	0.03

Node numbers correspond to the nodes in Figure 2, where ID_REF indicates the Affymetrix gene ID. Causal effect indicates the causal effect of each gene on the MMSE score.

Identification of gene with largest causal effect on MMSE scores

In order to identify the genes with the largest causal effect on MMSE scores, the causal effect of each gene on MMSE scores was estimated after creating a predictive model by LightGBM (Table 1). The gene expression changes of cell adhesion molecule 4 (*CADM4*) was found to have the largest causal effect on MMSE scores. As seen in Figure 1, the regression coefficient from *CADM4* (x2) to natriuretic peptide receptor 1 (*NPR1*) (x1) is 0.23, indicating that the increased expression of *CADM4* increases the expression of *NPR1*. Furthermore, the regression coefficient from *NPR1* (x1) to MMSE score (x0) is 0.18, indicating that increased expression of *NPR1* increases MMSE score. These results indicate that MMSE scores can be efficiently raised by increasing the expression level of *CADM4*.

Discussion

Altered gene expression of CADM4 affects gene expression of NRP1

CADM4 gene encodes cell adhesion molecule 4 (CADM4), which is expressed in oligodendrocytes in brain [9]. CADM4 is a member of CAMs (CADM1-CADM4), also known as synaptic cell adhesion molecules (SynCAMs), and regulates myelination and myelinated axon organization [10]. CADM4 binds to axonal CADM2 and CADM3 in the central nervous system (CNS) [11]. However, the function of CADM4 is still poorly understood, because oligodendrocytes can form normal myelin in mice lacking the *CADM4* gene [12].

NRP1 gene encodes natriuretic peptide receptor 1 (NPR1), which is mainly expressed in hippocampus and basal ganglia in brain [13]. NRP1 is a membrane-bound guanylate cyclase that serves as a receptor for both atrial natriuretic peptide (ANP) and brain natriuretic peptide (BNP) [14]. NPR1, ANP, and BNP are widely distributed in the nervous system and have functions in regulating neurotransmitter release and uptake, and in regulating signal transmission between synapses [15].

Figure 2 shows that the gene expression changes of *CADM4* influence *NRP1* with the coefficient 0.23, which represents an important causal relation because *CADM4* is one source of increasing *NRP1*. Since there are no reports of CADM4 directly affecting NPR1, further studies should be conducted on the relationship between CADM4 and NPR1.

Altered gene expression of NRP1 affects MMSE scores

ANP and BNP and their receptor, NPR1, are abundant in the CNS and have been shown to be involved in synaptic transmission, synaptic plasticity, neurovascular and blood-brain barrier integrity, inflammation, neuroprotection, and regulation of the hypothalamic-pituitary-adrenal (HPA) axis [16]. It was reported that the ANP-NPR1 pathway is neuroprotective against N-methyl-D-aspartate-induced neurotoxicity by activating dopamine D1 receptors [17]. Furthermore, increased plasma ANP and BNP may decrease the expression level of NPR1 in the CNS [18]. These reports suggest that increasing NPR1 expression improves MMSE scores, which is consistent with the present results that changes in NPR1 expression affect MMSE scores with a regression coefficient of 0.18 (Figure 2).

Conclusions

In this study, after extracting important genes contributing to the development of MCI using a random forest classifier, 26 causal genes that alter the MMSE scores by causal discovery were identified. In addition, *CADM4* was identified as the gene with the largest causal effect on MMSE scores. In this model, increasing the expression of *CADM4* can lead to improved MMSE scores.

Conflicts of interest

No competing interests to declare.

References

1. Saez-Atienzar S, Masliah E. Author Correction: Cellular senescence and Alzheimer disease: the egg and the chicken scenario. Nat Rev Neurosci. 2020;21(10):587.
2. Gao S, Hendrie HC, Hall KS, Hui S. The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. Arch Gen Psychiatry. 1998;55(9):809-815.
3. Manly JJ, Tang MX, Schupf N, Stern Y, Vonsattel JP, Mayeux R. Frequency and course of mild cognitive impairment in a multiethnic community. Ann Neurol. 2008;63(4):494-506.
4. Shimizu S, Hoyer PO, Hyvarinen A, Kerminen A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. J Mach Learn Res. 2006; 7(72):2003–2030.
5. Kotoku J, Oyama A, Kitazumi K, et al. Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups. PLoS One. 2020;15(12):e0243229.
6. Tsai JC, Chen CW, Chu H, et al. Comparing the Sensitivity, Specificity, and Predictive Values of the Montreal Cognitive Assessment and Mini-Mental State Examination When Screening People for Mild Cognitive Impairment and Dementia in Chinese Population. Arch Psychiatr Nurs. 2016;30(4):486-91.
7. Menze BH, Kelm BM, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009; 10:213.
8. Shimizu S, Inazumi T, Sogawa Y. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. J Mach Learn Res. 2011;12:1225–1248.
9. Elazar N, Vainshtein A, Rechav K, Tsoory M, Eshed-Eisenbach Y, Peles E. Coordinated internodal and paranodal adhesion controls accurate myelination by oligodendrocytes. J Cell Biol. 2019;218(9):2887-2895.
10. Elazar N, Vainshtein A, Golan N, et al. Axoglial Adhesion by Cadm4 Regulates CNS Myelination. Neuron. 2019;101(2):224-231.e5.
11. Sharma K, Schmitt S, Bergner CG, et al. Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci. 2015;18(12):1819-31.
12. Zhu Y, Li H, Li K, et al. Necl-4/SynCAM-4 is expressed in myelinating oligodendrocytes but not required for axonal myelination. PLoS One. 2013; 8(5):e64264.
13. Mahinrad S, Bulk M, van der Velpen I, et al. Natriuretic Peptides in Post-mortem Brain Tissue and Cerebrospinal Fluid of Non-demented Humans and Alzheimer's Disease Patients. Front Neurosci. 2018;12:864.
14. Suzuki T, Yamazaki T, Yazaki Y. The role of the natriuretic

peptides in the cardiovascular system. Cardiovasc Res. 2001;51(3):489-94.

15. Cao LH, Yang XL. Natriuretic peptides and their receptors in the central nervous system. Prog Neurobiol. 2008;84(3):234-48.
16. Hodes A, Lichtstein D. Natriuretic hormones in brain function. Front Endocrinol (Lausanne). 2014;5:201.
17. Kuribayashi K, Kitaoka Y, Kumai T, et al. Neuroprotective effect of atrial natriuretic peptide against NMDA-induced neurotoxicity in the rat retina. Brain Res. 2006;1071(1):34-41.
18. Mahinrad S, de Craen AJM, Yasar S, van Heemst D, Sabayan B. Natriuretic peptides in the central nervous system: Novel targets for cognitive impairment. Neurosci Biobehav Rev. 2016;68:148-156.