INCREASING THE ACCURACY OF THE DIAGNOSIS OF MENTAL DISORDERS BASED ON HETEROGENEOUS DISTRIBUTED DATA

A. Bogdanov[1], A. Degtyarev[1,2,a], N. Zalutskaya[3 c], N. Gomzyakova[3], S. Belavin[1], A. Khokhryakova[1]

[1] *Saint Petersburg State University, 7-9 Universitetskaya emb., Saint Petersburg, 199034, Russia*

[2] *Plekhanov Russian University of Economics, 36 Stremyanny lane, Moscow, 117997, Russia*

[3] *V.M.Bekhterev National Research Medical Center for Psychiatry and Neurology, 3 Bekhterev str., Saint Petersburg, 192019, Russia*

E-mail: [a] a.degtyarev@spbu.ru

There is no single diagnostic marker for neurodegenerative diseases. The biomedical data obtained during these studies have heterogeneous nature, which greatly complicates their collection, storage and complex analysis. Special methods of statistical analysis due to the described specifics of the data must be applied. The results obtained indicate that for a correct diagnosis, it is necessary to use a comprehensive assessment of all tests.

Keywords: mental disorders, Big data, clustering, diagnostic criteria

Alexander Bogdanov, Alexander Degtyarev, Natalia Zalutskaya,
Natalia Gomzyakova, Sergey Belavin, Anastasia Khokhryakova

1. Introduction

Mental disorders of a neurodegenerative nature are one of the most pressing medical, economic and social problems today. The most common cause of neurocognitive disorders in elderly patients is Alzheimer's disease (AD) [1]. AD is an incurable, steadily progressive heterogeneous disease that leads to permanent disability of the patient, which requires constant care from relatives and medical personnel.

The causes of the development of the disease remain not fully understood, which determines the need to search for new diagnostic methods, develop algorithms and subsequent data analysis. Dynamics of the development of the disease is very individual, but a timely diagnosis provides an opportunity for an early start of treatment, which can significantly slow down the progression of the patient's disease and improve the social and labor prognosis. In this regard, the verification of early differential diagnostic criteria acquires special scientific and practical significance.

2. Features of biomedical data

There is no single diagnostic marker for neurodegenerative diseases. It is not known in advance how much information will be sufficient for making a diagnosis and which of the diagnostic tools will play a key role in this decision, since the procedure for determining it is not strictly formalized. This process is individual and often requires the collection of a big amount of information. Diagnostics consists of a comprehensive assessment of the clinical picture when doctor observes and interviews a patient, laboratory instrumental, experimental psychological research and neuroimaging data. In particular, neuropsychological examination [2], magnetic resonance imaging (MRI) [3], electroencephalography (EEG) [4], blood tests [5], genotyping [6] are used in the field of brain pathologies research. The biomedical data obtained during these studies have heterogeneous nature, which greatly complicates their collection, storage and complex analysis. At the same time, the collected and processed data may have a different volume with comparable significance. The data is stored and presented in different formats, even for the same kind of survey. As a result of patients study, several groups of data appear from various measuring devices (EEG, MRI, etc.), which have different formats and are processed by different software. Some of the data may have gaps for various reasons, including those depending on the patient. Medical information is strictly confidential and must be anonymized for further mathematical and statistical analysis. Due to the volume of research required for one person, its high cost, collecting a huge amount of data from such studies is very labor intensive.

In any case, in accordance with the CAP theorem and the introduced classification [7], the data under consideration belongs to the one of Big Data types. It combines diversity, incompleteness and lack of clear models in the subject area. Therefore, we need to use appropriate working methods.

3. Features of biomedical data processing

Modern methods of data analysis can help to solve the problem of sharing heterogeneous clinical and biological sources in brain research. However, with such processing, specific problems arise that must be considered in the process of work:

1. The need to consolidate data due to their heterogeneity and distribution
2. The data must be anonymized due to administrative requirements for the protection of personal information
3. Special methods of statistical analysis due to the described specifics of the data must be applied.

The nature of biomedical research is characterized by the work with poorly formalized information and the work with simple or heuristic models. Therefore, the collected information is

required to distinguish certain classes. In this case, it is initially important to separate the class of healthy patients from the classes of sick patients.

4. Processing results

Attempts to cluster using available data for only one type of tests proved to be insufficiently effective. Despite certain successes [8,9] achieved based on MRI data processing, further improvement in the accuracy of classifying and making a diagnosis faced problems with a limited sample data. To this end, it was decided to involve the data of other biomedical studies of the same patients. For these purposes, the following data was used:

- Wechsler test data and blood tests
- Data of MRI examinations of the brain – preprocessing with the FreeSurfer software package
- EEG data – coherent analysis

It was decided by experts to distinguish primary and secondary features in the source data. The division of all patients into groups is carried out in several stages. The groups are preliminarily divided according to the main characteristics in each of the tests. At the second stage, the division into classes is refined by comparing the results and adding secondary features to the processing.

To divide into groups, the ISODATA[1] [10] algorithm was applied. Its advantage is that it does not require preliminary analysis of input data and a priori setting of the number of clusters. ISODATA can dynamically change the number of clusters into which the data is divided, depending on the features that characterize each element of the sample. As a result, we get the optimal number of clusters for the given parameters. A blood test was chosen for the study in terms of determining the number of resistant classes of patients. Cholesterol, triglyceride, glucose, HDL and LDL were selected as the primary signs on the blood test. They were used to compare the elements. The secondary signs were gender, age, educational level, use of psychoactive substances, traumatic brain injury and hypertension. By dividing patients into clusters so that in each cluster there are patients with the most similar, and patients from different clusters with the most dissimilar indicators. Such a breakthrough will help to identify the characteristic features of a particular disease. According to secondary signs, it is possible to determine risk groups and prescribe an additional examination for the purpose of early detection of the disease. Initially, 5 clusters were set, but during the work this number decreased by 1. As a result, 4 groups of patients were obtained. In one of the groups, patients from the control group were absent, in the other, such patients were in the majority. This suggests that there are some signs that help distinguish a healthy person from a sick person. It should be noted that such a division added confidence in the expert assessment of specialists in the presence of exactly four groups.

Next, machine learning methods were applied to classify the data. In the pictures the classification results for each of the studies are presented: the Wechsler test, MRI and EEG. To separate sick patients from the control group of healthy people, the results were pooled taking into account additional features. As a result, the best result was obtained by selecting the characteristics given in table (significance of differences between clusters). In this case, it was possible to reduce the result to three clusters, in one of which the control group is completely absent, and in the other healthy people prevail over patients with dementia and Alzheimer's disease.

[1] ISODATA – Iterative Self-Organizing Data Analysis Techniques

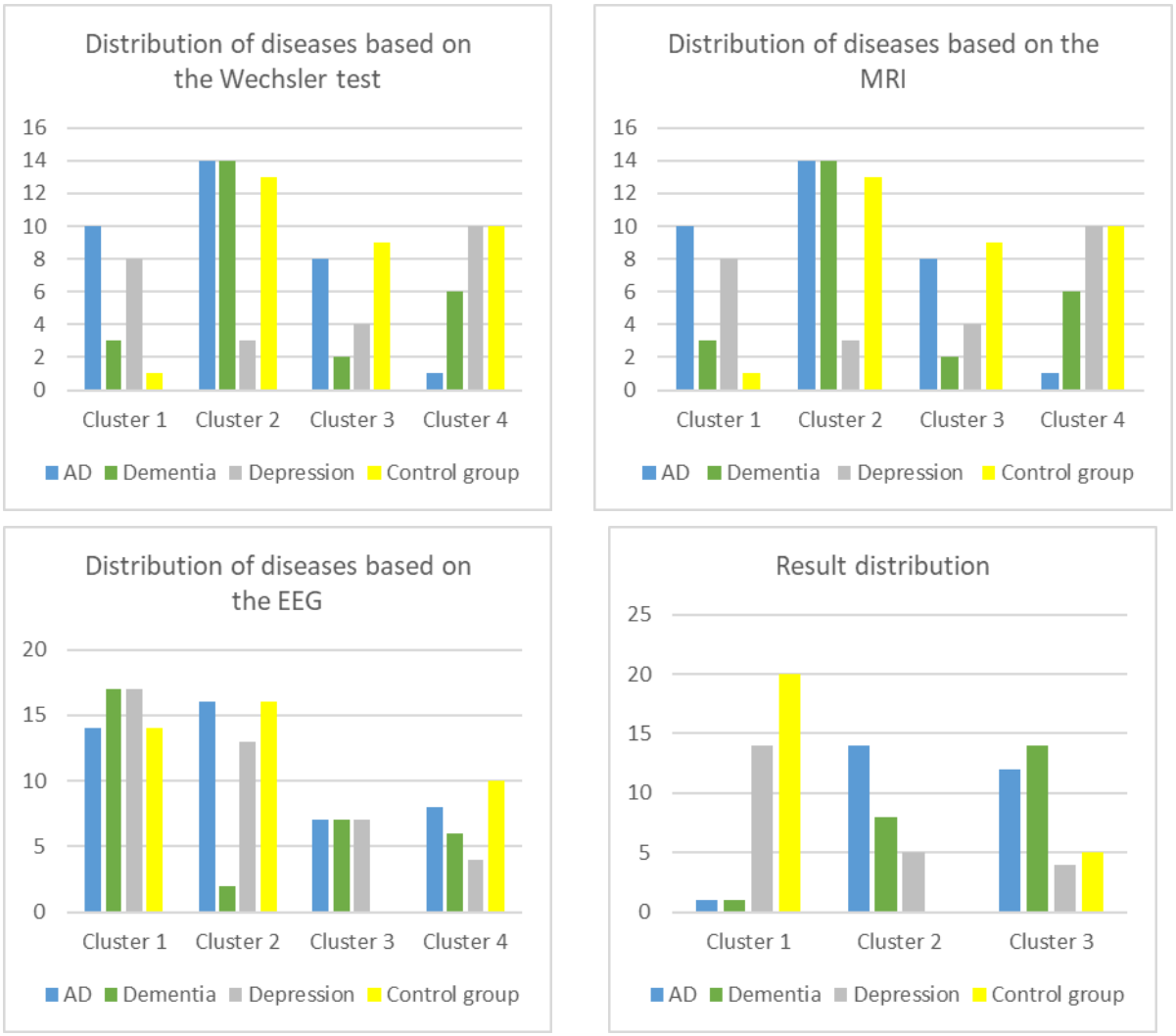

Figure 1. Classification results

Table 1. Significance of differences between clusters in accordance with Dan test

Indicator\Cluster	1-2	1-3	2-3	Indicator\Cluster	1-2	1-3	2-3
Right caudalmiddlefrontal ThickAvg	0,730	0,001	0,004	Right lateralorbitofrontal ThickAvg	0,707	0,175	0,150
Left parahippocampal GrayVol	0,000	0,000	0,523	Right lateraloccipital SurfArea	0,003	0,067	0,206
Left-Hippocampus	0,000	0,000	0,246	C4-Av Alfa	0,001	0,096	0,000
wm-rh-lateralorbitofrontal	0,002	0,284	0,032	F7-Av Alfa	0,091	0,000	0,000
Right lateraloccipital GrayVol	0,001	0,000	0,591	F8-Av Alfa	0,055	0,001	0,000
Left cuneus ThickAvg	0,484	0,114	0,399	T6-Av Alfa		0,004	0,000
Left frontalpole SurfArea	0,182	0,508	0,508	T3-Av Alfa	0,000	0,012	0,000
Left parsopercularis ThickAvg	0,145	0,556	0,061	T5-Av Alfa	0,008	0,019	0,000
Right lateraloccipital ThickAvg	0,024	0,024	0,917	F3-Av Alfa	0,119	0,000	0,000
Left-Accumbens-area	0,297	0,096	0,559	T4-Av Delta	0,035	0,072	0,000
Right entorhinal SurfArea	0,021	0,165	0,277	O2-Av Delta	0,998	0,003	0,004
Right parsorbitalis ThickAvg	0,080	0,171	0,003	B III	0,004	0,031	0,329
wm-lh-parsorbitalis	0,020	0,000	0,230	B Va right	0,855	0,955	0,855
WM-hypointensities	0,006	0,000	0,225	B VIIb diff	0,000	0,036	0,036
Left parahippocampal ThickAvg	0,000	0,001	0,512	Right isthmuscingulate ThickAvg	0,095	0,391	0,391

5. Conclusion

When considering each test separately, it is not possible to isolate biomarkers that allow a clear division of patients into groups.

Clustering according to selected features of the Wechsler test, MRI and EEG allows us to identify a group in which healthy patients were absent, which suggests that these features significantly distinguish the group of healthy patients and those with cognitive impairments. The results obtained indicate that for a correct diagnosis, it is necessary to use a comprehensive assessment of all tests.

The chosen method for analyzing data from examinations of the brain of patients with cognitive impairments and the results obtained give grounds to assert that research in this area is promising and requires further continuation, and it would be advisable to increase the number of tests under consideration.

References

[1] Budson A.E., Kowall N.W. Handbook of Alzheimer's disease and other dementias. Wiley-Blackwell, 2013; 387 p.

[2] Belleville S., Fouquet C., Hudon C. et al. Neuropsychological Measures that Predict Progression from Mild Cognitive Impairment to Alzheimer"s type dementia in Older Adults: a Systematic Review and Meta-Analysis. Neuropsychol Rev. 2017; 27: pp.328–353

[3] De Flores R., La Joie R., Chételat G. Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. Neuroscience. 2015; 309: pp. 29–50

[4] Nardone R., Sebastianelli L., Versace V., Saltuari L., Lochner P., Frey V., Golaszewski S, Brigo F, Trinka E, Holler Y. Usefulness of EEG Techniques in Distinguishing Frontotemporal Dementia from Alzheimer's Disease and Other Dementias. Dis Markers. 2018 Sep 3;2018:6581490

[5] Anstey K.J., Ashby-Mitchell K., Peters R. Updating the Evidence on the Association between Serum Cholesterol and Risk of Late-Life Dementia: Review and Meta-Analysis. J Alzheimers Dis. 2017; 56(1): pp. 215–228

[6] Neurodegenerative diseases: from the genome to the whole organism. Ed. by M.V.Ugrymov, Moscow: Nauchni mir, 2014, 848 p.

[7] Bogdanov A., Degtyarev A., Korkhov V., Thurein K., Shchegoleva N. Big data as the future of information technology. CEUR Workshop Proceedings, 2018, 2267, pp.26-31

[8] Korkhov V., at all Data storage, processing and analysis system to support brain research. LNCS, 2018, 10963, pp. 78-90.

[9] Volosnikov V., at all Data consolidation and analysis system for brain research. CEUR Workshop Proceedings, 2018, 2267, pp. 388-392

[10] Tou J.T., Gonzalez R.C. Pattern Recognition Principles. Addison-Wesley PC, 1974, 378p.