A Promising Approach to Detect Early Signs of Disease in Old Age

A. Bogdanov[1], N. Shchegoleva[1(✉)], and N. Zalutskaya[2]

[1] St. Petersburg University, Saint Petersburg, Russia
n.shchegoleva@spbu.ru

[2] Federal State Budgetary Institution «Bekhterev National Medical Research Psychiatry and Neurology Center», Ministry of Health of the Russian Federation, St. Petersburg, Russia

Abstract. The article proposes a platform structure that allows correctly, with a sufficient level of statistical significance, to separate a group of patients with the early stages of Alzheimer's disease from healthy controls, patients with late-life depression and people with cognitive impairment of vascular etiology. At the same time, three serious technical problems have been solved: the possibility of secure storage, transmission and provision of data in a form suitable for intellectual processing without disclosing personal information about patients has been implemented; filling in missing data to avoid distortion of survey results statistics; perform joint processing of data obtained at each of the diagnostic installations.

Keywords: Alzheimer's disease · protection of personal information · depersonalization · synthetic data sets

1 Introduction

Neurodegenerative diseases of the brain, accompanied by the development of dementia, represent a wide range of disorders, the common clinical sign of which can be considered a gradual decline in cognitive functions. Around the world, a diagnosis of Alzheimer's disease (AD) appears every 33 s and almost a million newly diagnosed cases are registered every year [1].

Despite the fact that the histopathological picture of AD is relatively well studied, for a long time intravital assessment of specific changes in brain tissue remained extremely difficult due to the lack of technical capabilities for their study and verification. That is why the first clinical criteria for diagnosing asthma were based on an analysis of the clinical picture and the nature of the disease. This has led to a wide range of conditions with different clinical manifestations being classified as Alzheimer's disease. Despite the possibility of a three-level assessment of the degree of probability of the presence of AD within the framework of this approach, however, to make a final diagnosis, a post-mortem histopathological examination of brain tissue was required. Thus, the diagnosis of "Alzheimer's disease" could only be established in the presence of dementia syndrome - a severe, irreversible, progressive and disabling condition, the treatment of which can pursue only very modest goals.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14821, pp. 358–368, 2024.
https://doi.org/10.1007/978-3-031-65308-7_25

The accumulation of information about the failure of attempts to restore the cognitive functioning of individuals whose cognitive decline has reached the level of dementia, as well as the emergence of technical capabilities for intravital study of functional and structural changes in brain tissue have become the basis for the development of new diagnostic [2].

At the same time, more than 100 years of experience in evidence-based research has demonstrated the futility of attempts to establish one single cause of the development of asthma, and, consequently, the identification of any one parameter as an indicator of the disease. Not a single sign detected by the above research methods is strictly specific for AD.

The typical onset of AD is a decrease in cognitive functions, primarily impairment of short-term memory, but atypical development of AD is possible. The early signs of Alzheimer's disease are similar to those of other age-related neurodegenerative diseases and early dementia, making early diagnosis of Alzheimer's disease extremely difficult [3].

Due to the extreme complexity of the development of the disease, a number of studies are performed to make a diagnosis, resulting in the formation of several groups of data, including clinical observation, anamnesis, results of experimental psychological tests, as well as data from various measuring instruments (for example, EEG, MRI and others), which have different formats and are processed by different programs and may be characterized by potential incompleteness, inconsistency, and noise. Such data is incompatible with each other without a comprehensive content analysis and non-trivial transformations to forms for which integral analysis is available.

At the same time, doctors can prescribe a different list of studies for each patient, and quite often the doctor's conclusion based on his experience may contradict the results of instrumental diagnostics.

Thus, we have the results of six (and often fewer) types of indirect diagnostics of patients, which are in no way related to each other, but their joint processing would make it possible to determine biomarkers that are associated with AD, which represents a classic problem of "Big Data" processing.

To store and analyze the results of medical examinations of patients at the Research Institute named after V. M. Bekhterev we are developing an information system that should ensure the protection of personal data, and in the future this system should be integrated with other brain research systems within the framework of international collaborations. All this leads to the need to solve three serious technical problems:

1. Possibility of secure storage, transmission and provision of data in a form suitable for intellectual processing without disclosing personal information about patients.
2. Filling in missing data to avoid distortion of statistics
3. Joint processing of data obtained on each of the diagnostic installations and correlation analysis, which allows you to calculate significant features.

2 Safe Handling of Data

The present time can most accurately be described by the word “digitalization.” Access to healthcare results in an ever-increasing amount of data, increasing the likelihood of data loss, corruption, and theft. Therefore, the importance of ensuring data security increases.

Data security has two aspects: the first is the personal data of patients, to which a limited circle of people should have access, and there is data that can be used to solve scientific problems and this is a matter of anonymization. The second is the data transmission medium. Using the Internet provides great opportunities, but does not provide data security, but using an intranet, where everything is protected, does not provide scientists with access to data sets.

A platform that can largely solve the above problems has been proposed by DGT Network (https://dgt.world/) and is called DGT [4]. DGT has a hierarchical topology in which each cluster of decentralized nodes - medical institutions - adapts to a specific business environment, while remaining part of the overall platform. DGT is a hybrid network, meaning that these clusters can have either public or consortium-based provisions to bring everyone together under one platform. DGT can process any digital value: IOT data, identifiers, digital twins, smart contracts, etc.

The complex structure of information exchange in this system is aimed at providing a single information space for different nodes, forming a single reliable copy of data, free from random and intentional errors. Such problems often arise when the system operates without a single center. The logic of checking each transaction and its distribution throughout the entire system (synchronization of registers) is described by a system of rules called consensus.

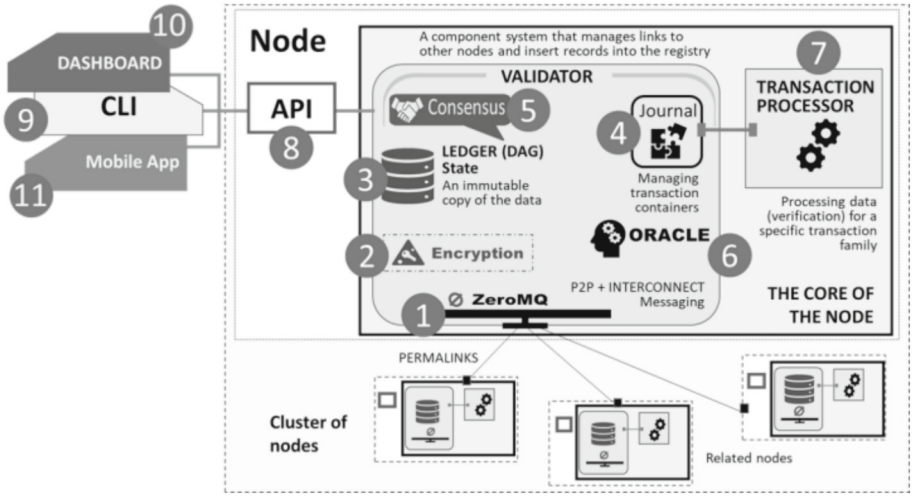

Fig. 1. Conceptual architecture of the DGT-based CFI platform.

The complex structure of information exchange in the described system is aimed at providing a single information space for different nodes, forming a single reliable copy of data, free from random and intentional errors. Such problems often arise when the system operates without a single center. The logic of checking each transaction and its distribution throughout the entire system (synchronization of registers) is described by a system of rules called consensus (see Fig. 1).

3 New Approach to Data Transfer

Recently, one of the most effective approaches is data protection in depth, that is, the use of several sequential measures to prevent access to information by third parties. This approach uses several levels of data protection, which increases system security and reduces the risks of unauthorized access to information. For secure data transfer, an algorithm has been developed for embedding encrypted personal information (full name, year and place of birth, profession, etc.), biometric information (weight, height, etc.), examination results, established diagnoses, etc.) in the form of a QR code into a color image, consisting of the following steps (see Fig. 1) [5]:

1. Preparing messages for QR codes with documentary information.
2. Message encryption for QR codes with documentary information
3. Highlighting the content of the barcode image
4. Dividing the selected content part of the barcode image into several parts
5. Application of geometric transformations of image matrices of barcode parts (rotation by an angle multiple of 90 degrees, matrix transposition, rotation around an axis that coincides with one of the rows or columns, etc.)
6. Write the converted parts of the barcode into different areas of the container image.
7. Accordingly, information extraction is carried out in reverse order.

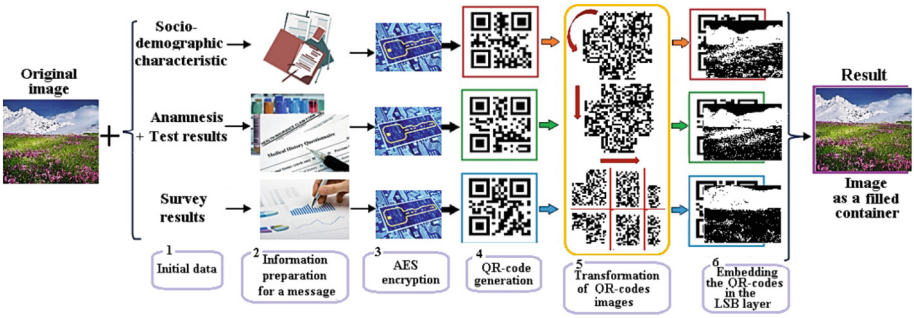

Fig. 2. New approach to data transfer and storing.

It should be noted that container images obtained in this way do not allow us to detect the existence of documentary information in them. The use of color images with embedded QR codes containing encrypted information as a storage and carrier of personal data will ensure the protection of this information, as well as limit public access to it, since

the embedded information can only be read and decrypted if there is an encryption key and knowledge of the structure of recording information.

This technology will make it possible to apply a universal standard when transferring personal data over the network, exchanging information, including through international channels, without losing any part of it and/or replacing it, as well as significantly reducing the losses of medical institutions from data leaks and significantly increasing security processing of personal data.

To effectively solve the problem of making a diagnosis, it is necessary to use all available patient examination data, while ensuring the protection of their personal data. This problem can be solved by anonymizing data.

4 Methods of Depersonalization

Depersonalization methods (classical depersonalization or k-Anonymity model) are aimed at eliminating the connection of individual attributes with identifiers [6]. The main idea for this class of methods is to identify groups of similar records (equivalence classes) that can be attributed to more than one individual. In this case, the main property of the definition of personal data is violated - the correlation of information with a directly or indirectly identified or identifiable individual. The size of this equivalence class (the size of similar records) allows us to calculate the probability of re-identification:

$$P_{datarisk} = \frac{k(D)}{\left\langle \widetilde{E} \right\rangle}$$

Here $\left\langle \widetilde{E} \right\rangle$ is the minimum size of all equivalence classes within the selected data set. If this size is determined by the number of rows in which all indirect attributes match, then this metric is called k-Anonymity. If additional statistical distributions are required in each equivalence class, more stringent metrics are used - ℓ-Diversity or τ-Closeness.

The advantages of this group of methods include their good determinism, resistance to binding attacks, and the ability to calculate metrics. On the other hand, applying the methods to large data sets requires a fair amount of computational effort.

Mondrian Multivariate Algorithm

The most suitable depersonalization method, in our opinion, is the Mondrian algorithm. This is a greedy multidimensional algorithm that recursively partitions the anonymized set into regions, each of which contains at least k-records. This uses the median split method on the attribute with the widest range of values.

The algorithm is named after the Dutch artist Piet Mondrian, whose abstract compositions inspired the geometric approach to sectioning used by the algorithm. The Mondrian algorithm works by partitioning data into rectangular subsets, or "bins," based on the values of attributes in the data set.

The Mondrian algorithm is an extremely useful anonymization tool and can be used for both batch and streaming data. At the same time, the method has some limitations. For example, the partitions generated by the algorithm may be highly dependent on the order of attributes and privacy settings, which may affect the quality of the anonymized data. Additionally, the algorithm can be computationally complex, especially for large data sets.

5 Synthetic Data

The studies have shown that it is necessary to fill in the missing values in the data, since during processing the computer perceives a zero value in the absence of data as an outlier, which is a gross error. The solution to this problem can be the generation of synthetic data.

Synthetic data are sets generated by the system to replace real values. Such sets represent data that is not correlated with real individuals, but retains the patterns of the original set. There are different ways to generate synthetic data. This study used an algorithmic approach.

Algorithmic synthetic data sets

In this case, artificial (synthetic) data is created in accordance with specified algorithms. The advantage of such sets is the predictable profile of the generated set. The disadvantage is the excessive rigidity of the algorithms, the description of which for each set requires additional research costs.

The algorithmic approach is used in situations where the raw data cannot be disclosed due to privacy issues, such as medical or financial data. There are several algorithmic methods for generating synthetic data, including:

1. *Random sampling:* This method involves random sampling from the original data set to create a synthetic data set with similar statistical properties. However, this method can result in biased or poor-quality data if the sample size is too small or if the original data has complex relationships.
2. *Model-based methods:* These methods use statistical models to create synthetic data that mimic the distribution of the original data. Examples of model-based methods include Gaussian mixture models, Bayesian networks, and decision trees.

Since the data has a huge number of gaps, the first step in solving the problem will be to fill them with synthetic data. First of all, the data were checked for normality distributions using various statistical criteria. Also for each analysis, a comparison was made between the mean values and variances between groups of men and women. For all signs graphs were constructed for visual assessment of the distribution of general totality. From the data obtained, a conclusion was drawn about the features, and each feature that was taken into account when filling in the gaps.

1. Filling with a constant value (mean, median, constant – whichever fits the data best): Missing values in a column were filled with statistical metrics such as mean or median, or constant a value preselected based on the context of the data.
2. Filling according to normal distribution: Missing values were filled with randomly selected values from the normal distribution, the parameters of which were determined either from the sample or taking into account reference tests values, if known. This approach allows for better simulate missing data and does not generate outliers.
3. Filling with a constant value with variation: Missing values are replaced with a constant value, but with adding variation to account for diversity in the data. This approach is an improvement on the first method, but it has not been used.
4. Iterative filling using sklearn.IterativeImputer: Missing values are predicted using the model RandomForestRegressor and this process is repeated several times for

clarification of the results. This method was used for features, for whose distribution or other knowledge about the character is unknown data and passes. One can try other models in the future (the sklearn package offers quite a few, but they all require careful selection of hyperparameters) and find a model that provides the best filling in terms of clustering missing data.

6 Collaborative Data Processing

It was noted above that all survey results obtained must be used in close connection, since the problems under consideration may have multi-criteria dependencies that are indistinguishable when using a single source of information. At the same time, the analysis should be carried out on entire sets of patient examination results using intelligent data processing methods based on deep machine learning and multivariate statistical analysis.

The analysis process should be based on complex techniques, the effective use of which requires easily accessible means of handling clinical and biological data, which, in turn, can be implemented through modern cloud and high-performance technologies.As a starting point in solving the problem, it is proposed to solve two problems that are classic for machine learning and multivariate statistical analysis - the classification problem and the clustering or cluster analysis problem. Both tasks are aimed at combining objects into groups according to characteristic common features and further identifying new objects into formed groups.

To achieve the goal of finding informative biomarkers of Alzheimer's disease, the classification task within the framework of the ongoing study is suitable due to the presence of clinical diagnoses. These physician diagnoses can act as predefined categories—classes—into which the patients in question can be placed. However, when solving this problem, it is worth taking into account the possibility of a mistake made by a specialist in the diagnosis made. Such cases are quite acceptable, taking into account all the previously described difficulties in diagnosing one or another type of neurodegenerative disease. The presence of errors in the diagnoses of specialists may not have the best effect on the final quality of the resulting solution. Therefore, it is critical that the diagnosis be correct, which is a significant limitation in the application of classification models. In this regard, applying classification algorithms in isolation may not be the most optimal solution. To achieve better accuracy in predicting classes, it is proposed to use not only clinical diagnoses established by specialists, but also diagnoses based on diagnostic classes obtained through data analysis using suitable methods. This approach to making diagnoses is called "data-driven". One suitable method for automatically identifying classes is cluster analysis.

To solve this problem, a software solution that implements the process of automatic classification of patients with the subsequent use of the resulting diagnostic labels to solve the classification problem. Solving both the clustering problem and the classification problem is equally important for the results of the study.

Classification methods are useful because in most cases they allow one to derive patterns and connections between data explicitly - in the form of functions or a list of conditional rules by which classification was made. Analysis of the obtained functions or rules will make it possible to clearly establish the patterns of interest or state their

absence, as well as discover new ones, if any. Because classification implies the presence of strictly defined categories of patients, among them there will be Alzheimer's patients. In the case of a high-quality trained classifier that can distinguish target patients from others with a sufficient level of accuracy, it will be possible to identify the rules by which the classifier identifies such patients, which can provide greater understanding of how to diagnose the disease. In addition, many classification models make it possible to evaluate the contribution of one or another object parameter (patient test values, be it EEG, MRI or blood test) on the classification result. This in turn can also give specialists a broader view of the nature of the disease.

Cluster analysis, as has been said many times before, allows you to find hidden patterns and new structures in data that would be impossible to detect by other methods. Cluster analysis is more difficult than classification in terms of extracting meaning from the results obtained, because does not provide an interpretation of why certain groups were formed. Interpretation of the resulting groups is already the task of specialists in the field of research. And this is the value of clustering methods - they give researchers food for thought, offer new hypotheses and new space for studying the topic. As part of the study, ideally, in the process of cluster analysis, a cluster consisting primarily of Alzheimer's patients should be identified. Identification of a separate group of such patients will allow for a more detailed comparison of patients in this category with the rest of the observed patients and, potentially, will bring more clarity to the issue of diagnosing

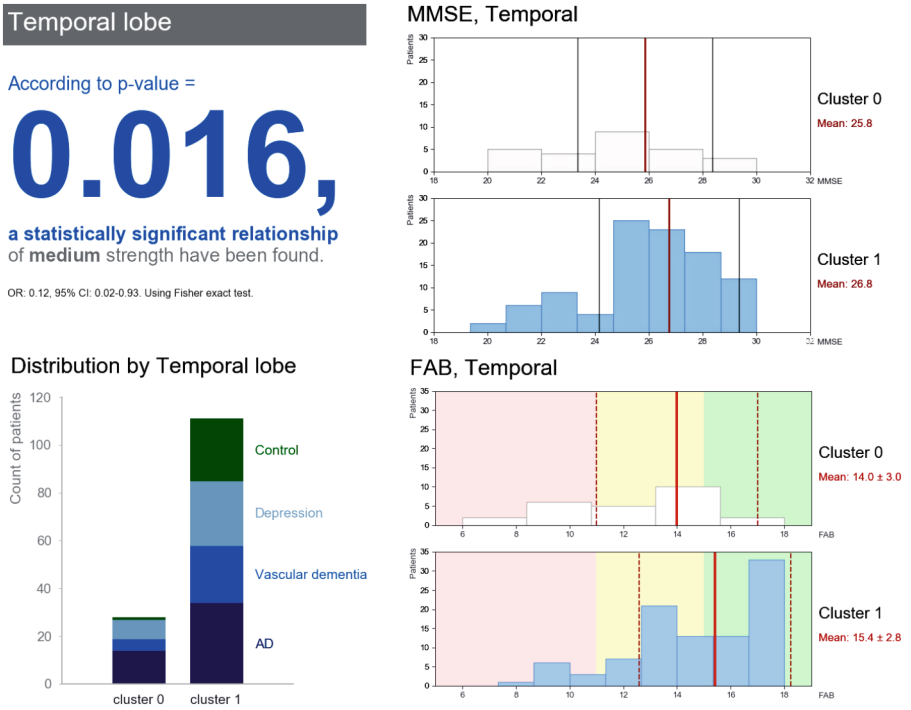

Fig. 3. Research results.

the disease. In addition, as in the case of classification, for clustering algorithms there are also methods that evaluate the contribution of a particular feature (parameter) of an object to the formation of a cluster. Knowing which feature contributed to the formation of the cluster to a greater extent can give specialists a deeper understanding of the nature of each group of patients and additional information for analyzing the formed categories. Together, all of the above should contribute to the discovery of new biomarkers of the disease.

The use of this approach to evaluate data from individuals with complaints of memory loss showed that cluster analysis allows one to correctly, with a sufficient level of statistical significance, separate the group of patients with the early stages of Alzheimer's disease from healthy controls, patients with late-life depression and individuals with cognitive impairment of vascular etiology (Fig. 2). The data we obtained as a result of multidimensional statistical processing allows us to talk about the possibility of correctly separating the indicators of individuals with neurodegenerative changes in the brain from those who are conditionally healthy (Fig. 3).

7 Conclusions

1. It is not enough to carry out measurements using only one, no matter how complex, diagnostic tool to identify illness of the elderly people.
2. To identify correlations, it is necessary to use data from at least three diagnostic installations.
3. To obtain reliable results, it is necessary to fill in all fields to perform statistical analysis in order to avoid distortion of the diagnosis.

The article proposes a platform structure that allows correctly, with a sufficient level of statistical significance, to separate a group of patients with the early stages of Alzheimer's disease from healthy controls, patients with late-life depression and people with cognitive impairment of vascular etiology. At the same time, three serious technical problems have been solved: the possibility of secure storage, transmission and provision of data in a form suitable for intellectual processing without disclosing personal information about patients has been implemented; filling in missing data to avoid distortion of survey results statistics; perform joint processing of data obtained at each of the diagnostic installations.

The data obtained as a result of multidimensional statistical processing allow us to talk about the possibility of correctly separating the indicators of individuals with neurodegenerative changes in the brain from those who are conditionally healthy.

This approach provides ample opportunities for further research in the field of psychiatry and neurology for the diagnosis and treatment of a wide range of mental and neurological disorders, using new categories of data and new highly effective data analysis methods for analysis.

To turn our research into practical medical technology it is necessary to carry out a number of computational experiments, each of which is specified by the following characteristics:

– Name and description of the dataset used for experiment, with details of its preprocessing

- Description of the feature extraction method
- Description of the clustering method
- Description of the resulting clusters
- Graphical presentation of results
- Calculation of quality assessment metrics
- Based on the results obtained, compare different combinations of methods and find the most preferred ones (by time, accuracy or other parameters)

Background Information

Personal data – information relating to a directly or indirectly identified or identifiable individual.

Processing of personal data - collection, recording, systematization, accumulation, storage, clarification (updating, changing), extraction, use, transfer (distribution, provision, access), depersonalization, blocking, deletion, destruction of personal data.

Personal data protection is a set of measures aimed at protecting information relating to a specific individual or determined on the basis of such information.

Violation of the security of personal data is divided into three categories: confidentiality (inaccessibility of information to unauthorized persons), integrity (preservation of completeness and consistency of information), availability (information is available and ready for use by authorized participants in information exchange).

Anonymization - the impossibility of re-identifying data, the irreversibility of depersonalization even when using additional information. Moreover, theoretically, if there are unlimited resources, any anonymous data can be re-identified.

Pseudonymization is the replacement of identifying information with a pseudonym (with the result that personal data can no longer be attributed to a specific data subject without the use of additional information, and such additional information is stored separately and is subject to technical and organizational measures to ensure confidentiality). Currently, the main attention should be paid to data exchange mechanisms (not only storage and reception). It is necessary to move to dynamic security based on changing parameters and scenarios, taking into account the context and characteristics of the data itself.

Depersonalization - actions as a result of which it becomes impossible, without the use of additional information, to determine the ownership of personal data to a specific subject of personal data.

Direct identifier is an attribute of personal data that allows you to identify the subject of personal data without involving additional information or other attributes. Example – passport number or TIN.

Quasi-identifier - an attribute of personal data that allows you to identify the subject of personal data together with other attributes - quasi-identifiers (that is, it is quite well correlated with the subject of personal data). Example – presence of a rare disease + address = unambiguous correspondence with a specific individual – subject of personal data.

An indirect identifier is an attribute of personal data that allows you to identify the subject of personal data only when additional information is involved.

References

1. Thies, W., Bleiler, L.: 2013 Alzheimer's disease facts and figures. Alzheimer's Dementia **9**(2), 208–245 (2013). https://doi.org/10.1016/j.jalz.2013.02.003
2. Dubois, B., et al.: Alzheimer disease: from biomarkers to diagnosis. Revue Neurologique **169**(10), 744–751 (2013). https://doi.org/10.1016/j.neurol.2013.07.016
3. Kulichikhin, K.Y., et al.: Development of molecular tools for diagnosis of Alzheimer's disease that are based on detection of amyloidogenic proteins. Prion T. 15. № 1. C. 56–69 (2021)
4. https://dgt.world
5. Shchegoleva, N., Zalutskaya, N., Dambaeva, A., Kiyamov, J., Dik, A.: New technologies for storing and transferring personal data. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) Computational Science and Its Applications – ICCSA 2022 Workshops. ICCSA 2022. Lecture Notes in Computer Science, vol. 13380. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10542-5_47
6. Bogdanov, A., Shchegoleva, N., Khvatov, V., Dik, G., Kiyamov, J., Dik, A.: K-anonymity versus PSI3 for depersonalization and security assessment of large data structures. In: Gervasi, O., et al. Computational Science and Its Applications – ICCSA 2023 Workshops. ICCSA 2023. Lecture Notes in Computer Science, vol. 14109. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37120-2_21