ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 4, No. 1, March, 2023

# Development and Algorithmization of a Method for Analyzing the Degree of Uniqueness of Personal Medical Data

Abas H. Lampezhev [1], Vladimir Zh. Kuklin [1], Leonid M. Chervyakov [1], Aslan A. Tatarkanov [1*]

[1] *Institute of Design and Technology Informatics of RAS, Russian Federation.*

Received 26 December 2022; Revised 21 February 2023; Accepted 25 February 2023; Published 01 March 2023

**Abstract**

The purpose of this investigation is to develop a method for quantitative assessment of the uniqueness of personal medical data (PMD) to improve their protection in medical information systems (MIS). The relevance of the goal is due to the fact that impersonal PMD can form unique combinations that are potentially of interest to intruders and threaten to reveal the patient's identity and medical confidentiality. Existing approaches were analyzed, and a new method for quantifying the degree of uniqueness of PMD was proposed. A weakness in existing approaches is the assumption that an attacker will use exact matching to identify people. The novelty of the method proposed in this paper lies in the fact that it is not limited to this hypothesis, although it has its limitations: it is not applicable to small samples. The developed method for determining the PMD uniqueness coefficient is based on the assumption of a multidimensional distribution of features, characterized by a covariance matrix, and a normal distribution, which provides the most reliable reflection of the existing relationships between features when analyzing large data samples. The results obtained in computational experiments show that efficiency is no worse than that of focus groups of specialized experts.

*Keywords:* Medical Information Systems; Personal Medical Data; Information Security; Medical Secret; Assessing Data Uniqueness.

## 1. Introduction

Currently, in all subject areas of human activity, studies aimed at improving data processing technologies are of great practical importance [1–4]. Data processing in MIS is no exception to this trend, but it has some fundamental features related to the information security of PMD. Based on the principles of system analysis, such data can be attributed to being unique. The problems associated with their description and solving related non-standard tasks can be eliminated using additional options, from data collection and analysis to data encryption and destruction, provided that the latter is necessary [5, 6].

Managing PMD is a complex and, in some ways, multidimensional problem. A particular management action requires a rational approach, considering the need for increased responsibility. The need to develop and implement new, more efficient methods of managing PMD processing is a characteristic feature of the modern process of intensive digitalization [7]. One of the critical tasks here comes down to ensuring an improved quality of healthcare services in commercial and public institutions.

Using unique PMD in various parts of MIS exacerbates the problem of ensuring their information security, i.e., protection against unauthorized access [8, 9]. At the head of the system of information security principles is maintaining the integrity of patients' medical data and ensuring their confidentiality and accessibility to the competent authorities on a legislative basis [10, 11].

* Corresponding author: as.tatarkanov@yandex.ru

http://dx.doi.org/10.28991/HIJ-2023-04-01-09

If an individual is unique in the population, then their risk of identification can be quite high. For example, individuals often cite privacy and confidentiality concerns and a lack of trust in researchers as reasons for not having their health information used for research purposes [12]. One of the factors that helps make the public more comfortable with their health information being used for research purposes is its de-identification at the earliest opportunity [13–15]. As many as 86% of respondents in one study were comfortable with the creation and use of a health database of de-identified information for research purposes, whereas only 35% were comfortable with such a database that included identifiable information [16].

A number of different uniqueness estimators have been proposed in the literature. It is important to know which of these works best for clinical data sets. One information protection mechanism proposed in the literature is differential privacy [17, 18]. Generally speaking, differential privacy requires that the answer to any query be "probabilistically indistinguishable" with or without a particular row in the database. Thus, differential privacy hides the presence of an individual in the database by making the two output distributions (with or without the row) "computationally indistinguishable" [19]. This is typically achieved by adding Laplace noise to every query's output.

However, for the context we are considering in this article, individual-level disclosure and differential privacy do not yet provide a ready-made solution, while uniqueness management has been the accepted approach to disclosure control over the past two decades. A weakness in existing approaches is the assumption that an attacker will use exact matching to identify people. The novelty of the method proposed in this paper lies in the fact that it is not limited to this hypothesis, although it has its limitations: it is not applicable to small samples.

In some cases, anonymized medical data can form unique combinations, representing potential interest to intruders and threatening to expose the patient's identity and medical secrets. Accordingly, it is an urgent task to develop a method for the early detection of unique data combinations for their subsequent additional protection during storage and processing in MIS. Accordingly, this study aims to develop a method to quantify the uniqueness of PMD to improve the processes of their protection, storage, and processing in MIS. The scientific novelty of the study is in developing a procedure for data processing based on the method of assessing their uniqueness, which makes it possible to clarify in an automated mode when detecting unique data combinations, the sequence of further actions with them, and to determine the conditions of access to information. The subsequent text describes the peculiarities of this procedure, which improves the system of medical care and access differentiation in specialized DBMS by preventing possible errors and abuses by users.

## 2. Literature Review

The essential element of developed information systems (IS), including medical ones as well as computer networks (CN), is the availability of specialized technologies and algorithms designed to regulate users' work with the target information in the interests of information security [19, 20]. The IS and CN system administrators must first reconcile with the institution administration the rights of each user to access the data. Figure 1 presents in more detail a scheme that describes in generalized form the structural aspects of the system that somehow delimit access to data. It is important to emphasize here that the distribution of access rights is a prerequisite for protecting computer systems.

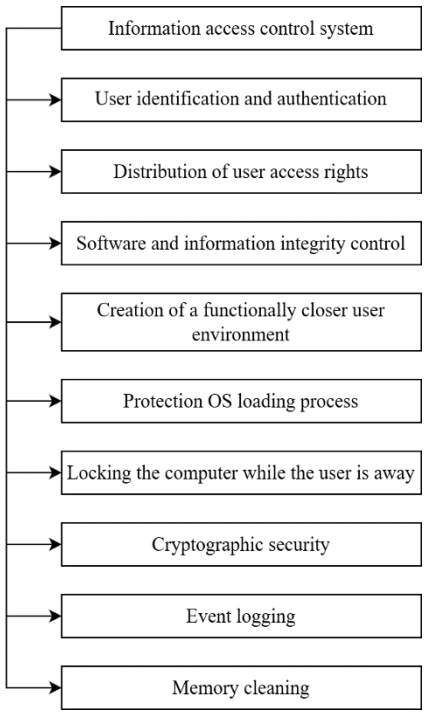


**Figure 1. Technologies used to differentiate data access**

Within MIS, all users are classified into different groups engaged in performing a specific list of functional tasks, including those based on such hardware and software as automated workstations: of a general practitioner, general practice nurse, specialist doctor, head of the department, chief physician, statistician, etc.; a wide range of MIS subsystems are also classified, which ensure the interaction between various medical organizations [21]. The capabilities of each user in the system are determined based on these tasks.

Among the used approaches, the role model of access control deserves attention [22, 23] because it is characterized by sufficient simplicity of administration and offers ample opportunities for setting security policies. However, within such a system, there are problems with providing individual access since it is not always possible to fit all users into the available roles. In particular, it is necessary to consider that in the framework of MIS, doctors have broad enough access to information about patients, and not always this access can be distinguished, as it is very difficult to determine what information a specialist will need for treating a patient [24]. At that, the problem of protection from information leakage, i.e., information security, remains.

Monitoring a wide range of information security threats has shown that one of the main defenses is enterprise security control [25]. It includes examining and regularly analyzing system logs, tracking system errors, monitoring the functioning of programs in use, and monitoring user actions. Administering implies the selection of security events recorded in logs, enabling and disabling events in security logging according to a set algorithm [26, 27].

The most common risks of information leakage in the medical field that occur in practice can be identified [28, 29]. Among them, first of all, the following should be noted:

- Medical staff's discussion of patients' health conditions with citizens who have no official authority in the matter;
- Private or official correspondence of doctors, which includes personal data about patients, using unprotected communication channels or information media;
- Providing inappropriate information about the diagnosis, discussing the course of the disease with the patient's visitors in the walls of the treatment and preventive care institution;
- Obtaining information about seeking medical help;
- Leaving the workplace with confidential patient data without proper control by medical facility staff;
- Entering a password into a computer in the presence of a colleague or third parties, including patients;
- Sharing a computer with several employees of a medical institution under a common password;
- Accidental or intentional verbal leakage of PMD in the wards in front of unauthorized persons or during a telephone conversation;
- Recovery, logically, by a health care professional of new, previously unavailable limited information about a patient;
- Errors made by staff due to lack of computer literacy.

A quantitative assessment, obtained using expert methods, of the potential risk associated with a systematic or one-time violation of medical secrecy plays a critical role in minimizing risks [30, 31]. Such indicators help to objectively assess the danger of potential threats, vulnerabilities, size of the damage, and sources of threats and calculate one integral indicator for the entire protection system.

Note that there are still no unified approaches to assessing the indicators that characterize information security, fully or indirectly. As a vivid example that deserves attention, it is possible to consider the assessment of the security criteria of medical information characterizing its confidentiality, the essence of which lies in the step-by-step implementation of the following steps:

- Systematic formation of a list of tasks aimed at ensuring confidentiality based on the constructed goal tree;
- Specification of threats that predetermine the goal of the *i*-th information task solution;
- Determining the information security indicator $П_{\text{IS}}^{(i)}$ , which characterizes the risk of the threat associated with the *i*-th task;
- Determining the integral value of confidential information:

$$\sum_{i=1}^{n} \frac{П_{\text{IS}}^{(i)}}{n} \quad (1)$$

where *n* is the number of tasks

In the framework of this procedure and many others, the assessment of individual indicators is usually based on

expert analytical methods. Approaches to assessing general information security can use the principles of fields such as fuzzy logic and algorithmic methods of obtaining appropriate conclusions under conditionally a priori or given conditions. Approaches exclusively theoretical to the modeling of threats using conceptual, functional and mathematical models are known. Conceptual models define the structural aspects of a particular "environment." Also, thanks to them, it is possible to receive descriptions of the properties of elements and the relations occurring in the system using an informal language.

Among approaches to solving the problem of mathematical modeling of threats, procedures based on the concepts of graph theory to represent IS (CN) models as queuing systems, distributed computing systems, or directly in the form of graphs deserve attention [32, 33].

Theoretical and practical examples of formulation and implementation of a specific problem solution of threat modeling are sufficient (Table 1). Among these studies, note the work representing approaches to identify different features that contribute to recognizing dangerous situations for operating IS (CN) [34].

**Table 1. Examples of approaches to modeling risks and threats to information security**

| No | Existing approaches to modeling information security risks and threats |
|---|---|
| 1. | Identification of different features that contribute to the recognition of dangerous situations for operating IS (CN) |
| 2. | Analysis of the structural aspects of modeling and information security systems |
| 3. | Problems posed by the algorithmization of models to assess the degree of IS (CN) security |
| 4. | Analysis of the possibility of using conventionally typical models of risks and threats to information security concerning individual objects of IS (CN) |
| 5. | Building an analytical model designed to assess the efficiency of IS (CN) protection from potentially dangerous situations related to attempts of data distortion on carriers stored by objects and entities in the insurance sector |
| 6. | Formalization of the selection criteria for information security systems, considering the indicators of functioning IS (CN) and destabilization factors in the design of information security systems |
| 7. | Analysis of existing risks and threats to IS (CN) as well as methods to address them through strategic planning and so-called software tools to ensure awareness |
| 8. | Building a model based on the assumption that any decision in terms of information security can, in case of objective necessity, be adjusted and presented using the generator of protected information |
| 9. | Building a model of the whole set of negative scenarios in aspects of the functioning of a particular IS (CN) |
| 10. | Building a model of the influence of internal and external risks and threats to IS (CN), the primary purpose is personal data processing |
| 11. | Building a model of risks and threats to IS (CN) of the verbal type |

In addition to the works listed above, note [35] that considered the issues of assessing the data uniqueness in detail. In particular, it constructed a combined decision rule for quantitative assessment of the data uniqueness measure, which uses the following four previously known approaches in its work: Pitman's estimator [36, 37], Zayatz estimator [38], sliding negative binomial estimator, and mu-argus [39]. This decision rule selects an estimation method from the four named ones, depending on the size of the analyzed sample. Dankar et al. [35] indicate the severe limitations of the developed decision rule despite the fact that a positive effect is achieved. Re-identification risk indicators tend to deviate conservatively since the data sets contain different quality problems (duplicates, errors, etc.) and an attacker will use the methods for accurate reidentification. The hypothesis does not apply to small samples and is considered as its limitation; however, it does not limit the method proposed in the present research.

The main disadvantages, to varying degrees, inherent in the approaches to solving the problem of modeling risks and threats to information security listed in Table 1 are as follows:

- The increased attention to the parameters of a potential attacker against the background of refusing to consider their influence on the formation of risks and threats;
- The lack of consistency, i.e., within a particular model, both generalized and private data are presented simultaneously (without systematization);
- Subjectivity in terms of constructing lists of risks and threats due to the opinions of experts;
- The lack of separation of risks and threats to IS (CN) and data;
- The absence of an explicit description of the risks and threats to IS (CN) with attention to negative scenarios, the essence and content of which are not disclosed or described only superficially;
- The lack in most models of mathematical formalization (the description of each specific case comes down to phrases and words indicating some consequences).

In addition to Table 1, pay attention to the STRIDE threat model (STM) set of methods and techniques. This Microsoft development implements an approach to building secure systems, considering the aspects of modeling the negative scenarios represented by the probabilities of the realization of risk and threat. The model preparation in this context should be implemented at the design stage of a complex software solution. The approach is based on a classification scheme applicable to the full description of attacks, considering the types of vulnerabilities used to implement negative scenarios.

STM's list of potentially dangerous situations includes:

- Impersonation or data substitution;
- The fact of refusal to perform actions by a specific subject, provided that it is impossible to prove the opposite;
- Disclosure of information through access to it if it is objectively limited;
- Denial of service and others

STM can be considered as a variant of IS information security determination with confidentiality, integrity, and accessibility, which is a combination, counted as primary principles in software development. Thus, ways to ensure the information security of IS (CN) can rely on several, to some extent, mastered approaches to risk and threat modeling within conceptual, functional, and mathematical models of IS (CN), some of which are described in Table 2.

**Table 2. Examples of different models of a particular IS (CN)**

| Principle of the model operation | Features or disadvantages of the model |
|---|---|
| **The type of IS (CN) model – Queuing system** | |
| Particular risks and threats characterize the incoming requests to the system input. IS (CN) can have one of the following states:<br>• Risks and threats did not occur, and negative scenario implementation is absent;<br>• Risks and threats have occurred, but implementation is absent;<br>• Risks and threats have been implemented. | The model is a black box:<br>• A priori unknown what the IS (CN) objects are;<br>• Data concerning their interaction are absent. |
| **The type of IS (CN) model – Distributed computing system** | |
| The input is the address of the object carrying the message. The output is the result, represented as information concerning whether a particular message has been delivered.<br>At the physical level, the model establishes a really existing connection between the objects of the system.<br>At the data link layer, it comes down to determining the interaction of the hardware addresses of the "here and now" network adapters.<br>At the network level, the model establishes connections between objects in terms of so-called logic addresses. | All aspects of the interaction of the software part with the operating system for a particular case are not considered. |
| **The type of IS (CN) model – Graph** | |
| The essence of the model is that it connects aspects of the interaction of conditional areas of risks and threats and information protection systems.<br>For this purpose, a graph is applied in which the relations between risks, threats (set $T$), and objects of protection (set $O$) form the graph $\{T, O\}$. If you enter a set $M$ characterizing the protection area, the result will be a graph $S = \{T, M, O\}$. | Specific indications of the interactions of different objects on the protected "*field*" and their descriptions are absent, which does not make it possible to consider the graph model absolutely "productive" in practice. |

The conducted review of approaches to modeling IS (CN) makes it possible to argue that it is impossible to fully describe the objects of a particular system only with their help; the same holds for describing the aspects of the interaction between them. Here, in studies related to the formation of indicators of information security of data, a problematic issue, especially concerning unique PMD, is developing methods to quantify the level of data security, but given models and methods do not consider the data uniqueness, which can be a source of threat and the key to a breach of information security. Accordingly, an urgent task is to develop a method to identify combinations of unique data in advance for their subsequent additional protection during storage and processing in MIS.

## 3. Research Methodology

The degree of efficiency in solving the target problem and the danger of data leakage in IS significantly increases when it comes to unique information and its diverse processing, which are, respectively, the object and subject of this study. A striking example of such information is PMD. Research aimed at improving the technologies that ensure their

management in the context of processing is a complex task that requires a competent and productive approach in all aspects associated with applying the principles of system analysis and modeling as well as the use of the "tools" of mathematical statistics. Against this background, the total consideration is necessary that the increased responsibility in decision-making accompanying the management, providing it in one way or another, is essential.

Analysis of the uniqueness of medical information about the patient, which determines its conditional value, expressed in an increase in the probability of real goal achievement by those who have access to relevant data, makes it possible to identify attributes that characterize:

- Patients whose treatment technology differs from standard approaches;
- The need for careful quality control of the services provided;
- Rare diseases;
- Uncharacteristic development of the disease;
- Errors made during data registration;
- Errors made during research;
- Errors made during treatment;
- Facts of medical falsification;
- Individual peculiarities of the person;
- Threats to patient information security (it is about the probability of PMD leakage. It largely depends on the typicality or atypicality of the sets of relevant indicators).

Multidimensional analysis of technology development in the healthcare sector has shown that common solutions can only cope with conventionally standard tasks. They cannot analyze patient-related medical information while identifying datasets that stand out from general arrays and are not fully consistent with a typical dataset.

Correct evaluation and application of criteria such as the data uniqueness coefficient make it possible to systematize PMD processing and determine particular conditions for providing access (from general to situational). In implementing dynamic data access control, fundamentally new solutions can be found to reduce the influence of human factors and improve the quality of decision-making in automating the quality control procedures of provided real-time medical services. Figure 2 shows the features of interaction between several subsystems during the data processing procedure.

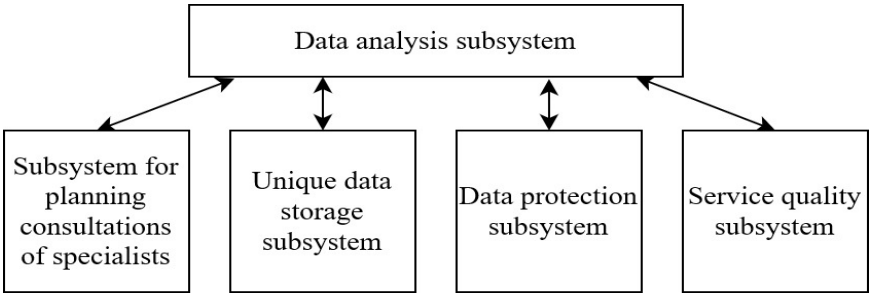


**Figure 2. Interaction of several information subsystems**

The most important part of the procedure for checking the possibility of access denial to unique data is the formation of a specific reference to the user passports generated as some database array storing all the necessary information about the users and their trustworthiness. Here, trustworthiness should derive from the characteristics and features of the person and social and individual conditions where the direct activity occurs. Figure 3 presents a flowchart of such process.

In the proposed model, the identification of unique medical data is implemented sequentially in the following stages:

- Formation of a vector of information and diagnostic attributes $\{X = X_1, X_2, ..., X_n\}$ corresponding to a certain patient;
- Calculation of the probability $P(X_1 = a_1, X_2 = a_2, ..., X_n = a_n)$ that the vector of information and diagnostic attributes reaches some threshold values characterizing the data uniqueness coefficient $K$ (see Figure 4):

$$P(X_1 = a_1\ X_2 = a_2\ \dots\ X_n = a_n) = \int_{\alpha_1}^{\beta_1}\int_{\alpha_2}^{\beta_2}\dots\int_{\alpha_n}^{\beta_n}\varphi(x_1\ x_2\ \dots x_n)\,dx_1 \dots dx_n \quad (2)$$

where $\varphi(x_1, x_2, ..., x_n)$is the density of the distribution of a random variable ($Xi, Xi,..., X$) under the condition of its distribution according to the $n$-dimensional normal law ($M$ – is the mean):

$$\varphi(x_1\ x_2\ \dots x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{|K_{ij}|}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{ij}^{-1}(x_i - M_i)(x_j - M_j)\right] \quad (3)$$

- Establishing the identification reliability threshold of the data set $\varepsilon$, with which the comparison of the probability $P(X_1=a_1, \dots, X_n=a_n)$ is then made: if this probability is less than $\varepsilon$, the data set is unique and can be associated with a particular individual.

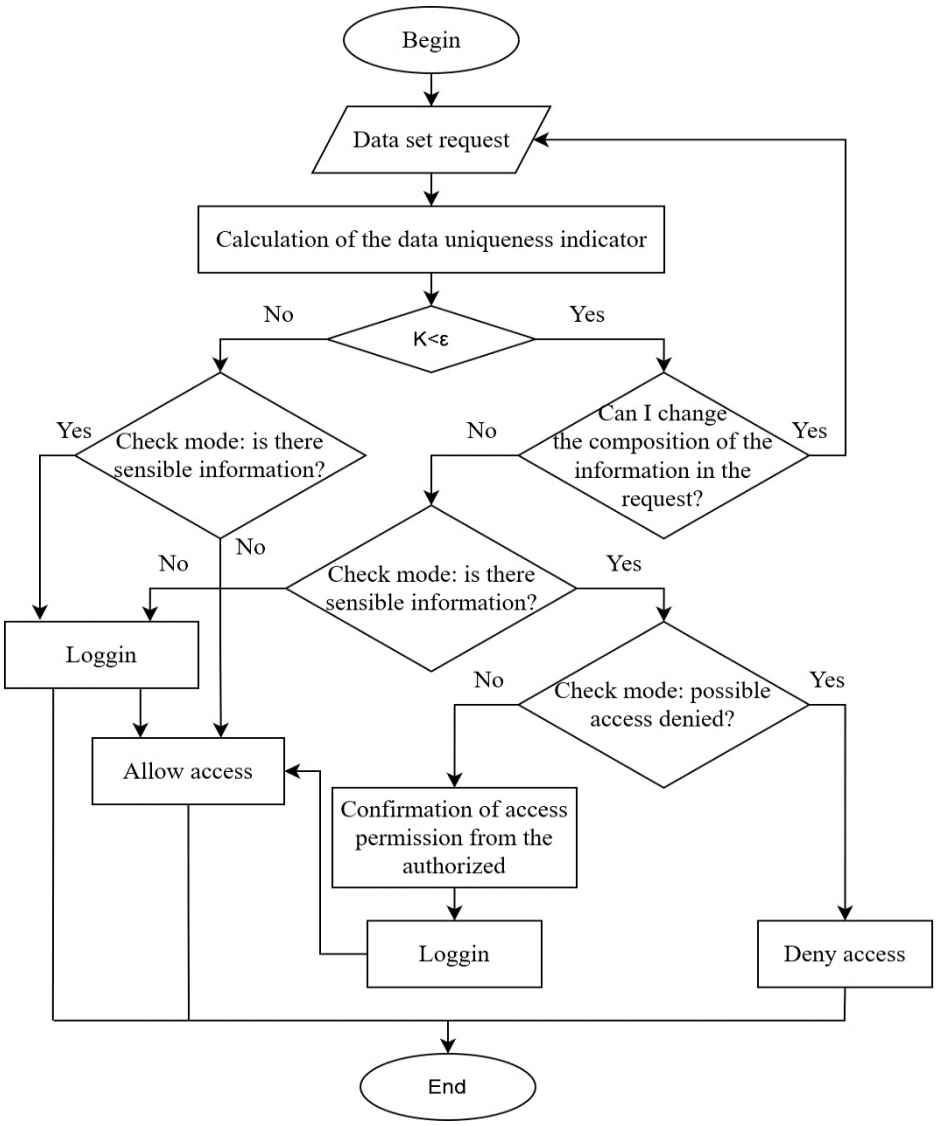


**Figure 3. Using a uniqueness coefficient to implement access control capability**

This methodological approach is based on the assumption of their multivariate distribution, characterized by the covariance matrix $K_{ij}$ of the normal distribution, which gives the most reliable reflection of the existing relationships between the attributes. Under certain conditions, distributions of any random variables tend to this law. The distribution law of accepted values of medical diagnostic parameters strives to the normal as the sample increases (long-term study) as well as increasing its representativeness and the division of initial attributes into groups of highly correlated attributes (within the group) according to their diseases (other attributes). Among other things, many diagnostic attributes are symmetrical, which is explainable by the standardization of the type or form and the way of identifying the medical indicators (below or above the norm).

To determine the previously noted value – threshold ε – it is possible to involve experts. In doing so, the so-called learning sample is additionally necessary, which will describe atypical cases in the form of rare combinations and sets of diagnostic criteria because of the analysis. The same applies to situations that cannot occur under actual conditions. Relevant information must be "transferred" and recorded in the class of unique data.

It is proposed to use cluster analysis methods when calculating the probability of taking specific values by set ($P$) to reduce the volume and labor intensity of calculation procedures in the marked model. The cluster analysis algorithms are numerous. They can be applied when working with different sets of criteria. Assessment of the clustering quality level $F(S)$ according to the proposed model is possible using the following dependencies:

$$F(S) = \sum_{l=1}^{k} \sum_{X_i X_j \in G_l} d(X_i\ X_j) + \sum_{l=1}^{k-1\Sigma} \sum_{m=l+1}^{k\Sigma} \frac{1}{d^p(\bar{X}(l)\ \bar{X}(m))} \rightarrow min \quad (4)$$

where $\bar{X}(l) = \frac{1}{n_l} \sum_{X_i \in G_l} X_i$ is the center of gravity (CG) of the group l; *d* is a proximity measure of objects; *k* are classes (number); *p* is the parameter (*p* = 1, 2, 3, ...) $\bar{X}(m) = \frac{1}{n_m} \sum_{X_i \in G_m} X_i$ is the CG of the group *m*.

$$F(S) = \sum_{l=1}^{k} \sum_{X_i \in G_l} d(X_i\ \bar{X}(l)) + \sum_{l=1}^{k-1\Sigma} \sum_{m=l+1}^{k\Sigma} \frac{1}{d^p(\bar{X}(l)\ \bar{X}(m))} \rightarrow min \quad (5)$$

$$F(S) = 1 - \sum_{l=1}^{k} \sum_{X_i \in G_l} d^2(X_i\ \bar{X}(l)) \,/\, \sum_{i=1}^{n} d^2(X_i\ \bar{X}) \rightarrow max \quad (6)$$

where$\bar{X}(l) = \frac{1}{n_l} \sum_i X_i$ is the CG of the set

$$F(S) = \sum_{l=1}^{k} \sum_{X_i X_j \in G_l} d(X_i\ X_j) \,/n + \sum_{l=1}^{k-1\Sigma} \sum_{m=l+1}^{k\Sigma} \frac{1}{d^p(\bar{X}(l)\ \bar{X}(m))} \rightarrow min \quad (7)$$

$$F(S) = \sum_{l=1}^{k} \sum_{X_i \in G_l} d(X_i\ \bar{X}_l) \,/n_l + \frac{1}{k} \sum_{l=1}^{k-1\Sigma} \sum_{m=l+1}^{k\Sigma} \frac{1}{d^p(\bar{X}(l)\ \bar{X}(m))} \rightarrow min \quad (8)$$

$$F(S) = \sum_{l=1}^{k} \sum_{X_i \in G_l} d(X_i\ \bar{X}(l)) \,/n_l + \sum_{l=1}^{k-1\Sigma} \sum_{m=l+1}^{k\Sigma} \frac{1}{d^p(\bar{X}(l)\ \bar{X}(m))} \rightarrow min \quad (9)$$

Here, the estimation of partitioning quality as a task is complex. The classification at the training stage implies repeated implementation under the condition of the replacement of metrics and parameters specified by the user. One of the reasonable options is to use several approaches immediately and compare the results.

## 4. Results and Discussion

According to the provisions outlined in the proposed model for managing the PMD processing, based on an assessment of their uniqueness, an algorithm was developed to analyze patient data security, providing for the implementation of procedures such as:

- Formation of a user's appeal to the system, aimed at gaining access to the data of a particular patient;
- Construction of correlation, covariance, and inverse (concerning the latter) matrices for set *P* with specific values;
- Clustering of objects;
- Identifying the correct, i.e., optimal method of "splitting" the set of "patient" type objects (and not only) into subgroups or classes under the condition of increased high correlation within one class and weak correlation between different classes;
- Determining the value of the index of the actual uniqueness of patient data used in forming the user's appeal.

Figure 4 shows a flowchart of the algorithm for analyzing the degree of the uniqueness of PMD, designed to assess patient data security.

Using the substantiated provisions of the developed PMD processing management model, based on assessing the degree of their uniqueness and considering information security requirements, an algorithm was developed to assess the degree of PMD uniqueness to analyze the level of patient data security. Determining the factual uniqueness coefficient for patient data is based on building correlation, covariance, and inverse matrices for the data set and subsequent clustering. This methodological approach assumes a multivariate distribution of features, characterized by the covariance matrix, and a normal distribution, which provides the most reliable reflection of the existing relationships between the features when analyzing samples of large-volume data.

To assess the degree of the reliability of the results obtained on the basis of the developed algorithm, a series of experiments were carried out. Samples of records from the database of patients, different in volume and composition, were formed. The sample size of records with repetition in each of the 10 experimental series conducted varied in the range from 150 to 240. The average value of the uniqueness coefficient for each series was calculated. Records with information about the values of medical indicators were provided to experts to highlight unique ones among them. The results of the expert assessment coincided with the results obtained using the developed method for analyzing the uniqueness of data. As unique, those are selected for which the values of the uniqueness coefficient are less than the average value by 12 times. The threshold value of uniqueness $\varepsilon = 1.4 \times 10^{-5}$. With an increase in the sample size, the threshold value of the coefficient increased. The results of the experiments showed that the values of the data uniqueness coefficients remained stable when the experimental conditions changed. A series of experiments were carried out to refine the threshold used to highlight unique datasets. During this experiment, experts received information about rare, in their opinion, combinations of values of diagnostic features. The coefficients of the uniqueness of these data were calculated when these sets were included in each of the above 10 series of experiments. The results of the experiment showed that the values of the coefficients of uniqueness of data received from experts remained stable: the deviation from the average value did not exceed 6%.

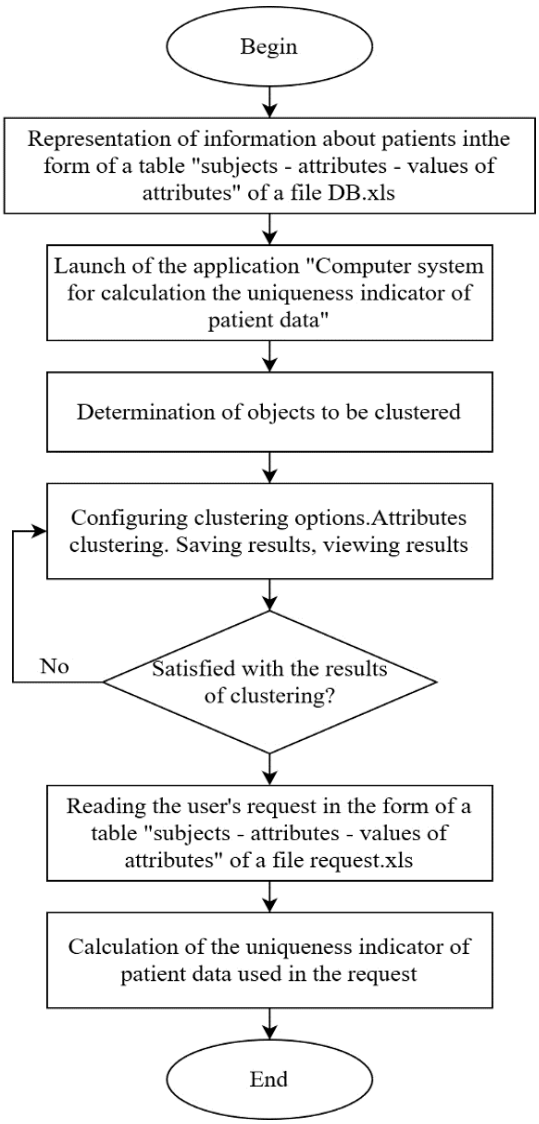


**Figure 4. Flowchart of the algorithm for analyzing the degree of uniqueness of PMD, designed to assess the patient data security**

Comparison of the results of this work with other studies shows that a common disadvantage of many studies is the special attention to the potential attacker parameters without considering their impact on forming the risks and threats. For example, despite the positive results achieved, the authors of Dankar et al. [35] note the severe limitation of the developed decision rule - one of the assumptions in their threat model is that an attacker will use accurate matching to reidentify people; however, the data sets contain errors, duplicates, and other quality problems. Thus, re-identification risk indicators tend to mistake the conservative direction. This hypothesis does not limit the method proposed in this paper (its strength), although it has another limitation: it does not apply to small samples.

When comparing the developed method with previously known methods, it should be noted that the weakness of the existing approaches is the assumption that an attacker will use exact matching to identify people. The novelty of the method proposed in this paper lies in the fact that it is not limited to this hypothesis, although it has its limitations: it is not applicable to small samples.

With regard to the investigation, it should be noted that the choice of the normal distribution law is justified by the fact that the distributions of both discrete and continuous random variables approach it under certain conditions. With an increase in the sample (long-term study), an increase in its representativeness, and the grouping of patients according to their diseases, the law of distribution of the accepted values of the analyzed parameters will tend to normalize due to the central limit theorem. Most of the values of diagnostic features are symmetrical due to the standard form of determining the values of medical indicators: below normal, normal, and above normal. The allocation of ranges of values for norms and pathologies suggests that most people should have (due to normal health or treatment) the values of diagnostic signs within the normal range. Patients with a common disease, due to homogeneity, should also have the values of the signs, on average, close to each other. It is known that many medical indicators (including those used in the study) have a normal distribution of their values.

In addition, the work carried out substantiates the need to apply the procedure for clustering diagnostic features to transform the integrand, which is due to the fact that the amount of computation during integration increases with the dimension of the integral, and the multiplicity of the integral can exceed several tens, which makes the existing computational methods practically inapplicable. Column vectors of the correlation matrix of features were used as clustering objects.

## 5. Conclusions

Based on the essence of MIS, both the need for effective use of PMD within it and its mandatory protection against unauthorized access are clear. If such data is unique, the problems associated with ensuring information security are significantly exacerbated. To solve these problems, a PMD processing management model was developed using various system analysis tools based on assessing the degree of their uniqueness and considering information security requirements. Its procedures identify (to analyze the level of patient data security) uncharacteristic combinations of data and quantify the degree of their uniqueness.

An algorithmic sequence of actions for assessing the degree of PMD uniqueness based on set P (which includes values of several diagnostic criteria) has also been proposed. The purpose of its use in practice is the categorization of patients and their assignment to "typical" and "atypical" groups.

The theoretical contribution of this study is that it proposes to use the clustering of parameters (diagnostic features) to simplify the computational process of calculating the uniqueness coefficient of patient medical data for further mathematical transformations of the primary expression used to estimate the data's uniqueness. Another peculiarity of the theoretical contribution is the study of the properties of the integrand to transform the integral by replacing variables, which reduces the size of the problem and the calculation volume when integrating; in particular, as is known, the labor intensity of the Monte Carlo method increases with the dimensionality of integrals. Nevertheless, the results of this work have limitations: it does not apply to small-size samples, although the above hypothesis does not limit the method proposed in this paper (its strength).

Computational experiments using the MIS simulation database and the developed algorithm analyzed the efficiency of calculating the data uniqueness coefficient for the selected group of patients. The obtained results showed that all patients with atypical combinations of data (considering expert opinions) were identified and detected using the developed algorithm with high accuracy (the deviation from the conditionally averaged value of the uniqueness coefficient was less than 6%).

The results obtained in the experiments give reason to expect the efficiency of the potential use of the proposed algorithm to provide an automated search of PMD secure from threats. The main practical recommendation for implementing the obtained results is their use in software development to identify unique or unreliable cases in medical practice, explained by patients' individual characteristics or medical errors (data falsification, errors in examining the patient). Developing such software on modular principles can be a direction for further work. Here, the final result of the software product may be:

- Forming a user request for patient data output;
- Forming correlation, covariance, and inverse covariance matrices for the set of diagnostic features and converting them to a specific form according to the above methodology underlying the calculation of the patient data uniqueness coefficient;
- Clustering of objects using the condensation search method based on the specified settings;
- Searching for optimal partitioning of a set of objects (patients, characteristics) into classes with a high correlation of objects within a group and a weak one between groups.

It is important to note that the main contribution of the obtained results is that they have no limitation to the hypothesis that an attacker would use accurate data matching to reidentify patients, especially given that in practice, datasets contain errors, duplicates, and other quality problems.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization, A.A.T.; methodology, A.A.T. and V.Zh.K.; software, A.H.L.; validation, A.A.T., and V.Zh.K; formal analysis, L.M.C.; investigation, A.A.T., A.H.L., and L.M.C.; resources, A.H.L.; data curation, A.A.T.; writing—original draft preparation, A.H.L. and L.M.C.; writing—review and editing, A.A.T. and V.Zh.K.; visualization, A.H.L.; supervision, A.A.T.; project administration, A.A.T.; funding acquisition, V.Zh.K. and A.A.T. All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available in the article.

### 6.3. Funding

Selected findings of this work were obtained under the Grant Agreement in the form of subsidies from the federal budget of the Russian Federation for state support for the establishment and development of world-class scientific centers performing R&D on scientific and technological development priorities dated April 20, 2022, No. 075-15-2022-307.

### 6.4. 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.

## 7. References

[1] Tatarkanov, A., Alexandrov, I., Muranov, A., & Lampezhev, A. (2022). Development of a Technique for the Spectral Description of Curves of Complex Shape for Problems of Object Classification. Emerging Science Journal, 6(6), 1455–1475. doi:10.28991/esj-2022-06-06-015.

[2] Lampezhev, A. H., Linskaya, E. Y., Tatarkanov, A. A., & Alexandrov, I. A. (2021). Cluster Data Analysis with a Fuzzy Equivalence Relation to Substantiate a Medical Diagnosis. Emerging Science Journal, 5(5), 688–699. doi:10.28991/esj-2021-01305.

[3] Tatarkanov, A. A., Alexandrov, I. A., Chervjakov, L. M., & Karlova, T. V. (2022). A Fuzzy Approach to the Synthesis of Cognitive Maps for Modeling Decision Making in Complex Systems. Emerging Science Journal, 6(2), 368–381. doi:10.28991/esj-2022-06-02-012.

[4] Tatarkanov, A., Alexandrov, I., & Glashev, R. (2021). Synthesis of neural network structure for the analysis of complex structured ocular fundus images. Journal of Applied Engineering Science, 19(2), 344–355. doi:10.5937/jaes0-31238.

[5] Chua, H. N., Ooi, J. S., & Herbland, A. (2021). The effects of different personal data categories on information privacy concern and disclosure. Computers & Security, 110, 102453. doi:10.1016/j.cose.2021.102453.

[6] Qamar, S. (2022). Healthcare data analysis by feature extraction and classification using deep learning with cloud based cyber security. Computers and Electrical Engineering, 104(A), 108406. doi:10.1016/j.compeleceng.2022.108406.

[7] Vitabile, S., Marks, M., Stojanovic, D., Pllana, S., Molina, J. M., Krzyszton, M., …, Salomie, I. (2019). Medical Data Processing and Analysis for Remote Health and Activities Monitoring. High-Performance Modelling and Simulation for Big Data Applications. Lecture Notes in Computer Science, 11400. Springer, Cham, Switzerland. doi:10.1007/978-3-030-16272-6_7.

[8] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M. (2017). Big data security and privacy in healthcare: A Review. Procedia Computer Science, 113, 73–80. doi:10.1016/j.procs.2017.08.292.

[9] Garcia-Perez, A., Cegarra-Navarro, J. G., Sallos, M. P., Martinez-Caro, E., & Chinnaswamy, A. (2023). Resilience in healthcare systems: Cyber security and digital transformation. Technovation, 121. doi:10.1016/j.technovation.2022.102583.

[10] Sharma, P., Borah, M. D., & Namasudra, S. (2021). Improving security of medical big data by using Blockchain technology. Computers & Electrical Engineering, 96, 107529. doi:10.1016/j.compeleceng.2021.107529.

[11] Kumar, R., Sharma, S., Vachhani, C., & Yadav, N. (2022). What changed in the cyber-security after COVID-19? Computers & Security, 120, 102821. doi:10.1016/j.cose.2022.102821.

[12] Nass, S. J., Levit, L. A., & Gostin, L. O. (2009). The HIPAA privacy rule. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. National Academies Press, Washington, United States. doi:10.17226/12458.

[13] Deng, Z., & Liu, S. (2017). Understanding consumer health information-seeking behavior from the perspective of the risk perception attitude framework and social support in mobile social media websites. International Journal of Medical Informatics, 105, 98–109. doi:10.1016/j.ijmedinf.2017.05.014.

[14] Willison, D. J., Swinton, M., Schwartz, L., Abelson, J., Charles, C., Northrup, D., Cheng, J., & Thabane, L. (2008). Alternatives to project-specific consent for access to personal information for health research: Insights from a public dialogue. BMC Medical Ethics, 9(1). doi:10.1186/1472-6939-9-18.

[15] Nair, K., Willison, D., Holbrook, A., & Keshavjee, K. (2004). Patients' consent preferences regarding the use of their health information for research purposes: A qualitative study. Journal of Health Services Research and Policy, 9(1), 22–27. doi:10.1258/135581904322716076.

[16] Kass, N. E., Natowicz, M. R., Hull, S. C., Faden, R. R., Plantinga, L., Gostin, L. O., & Slutsman, J. (2003). The use of medical records in research: What do patients want? Journal of Law, Medicine and Ethics, 31(3), 429–433. doi:10.1111/j.1748-720X.2003.tb00105.x.

[17] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, 3876. Springer, Berlin, Germany. doi:10.1007/11681878_14.

[18] Dwork, C. (2006). Differential Privacy. Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, 4052, Springer, Berlin, Germany. doi:10.1007/11787006_1.

[19] Wang, K., Xie, S., & Rodrigues, J. (2022). Medical data security of wearable tele-rehabilitation under internet of things. Internet of Things and Cyber-Physical Systems, 2, 1–11. doi:10.1016/j.iotcps.2022.02.001.

[20] Altameem, A., Kovtun, V., Al-Ma'aitah, M., Altameem, T., H, F., & Youssef, A. E. (2022). Patient's data privacy protection in medical healthcare transmission services using back propagation learning. Computers and Electrical Engineering, 102, 108087. doi:10.1016/j.compeleceng.2022.108087.

[21] Fruehwirt, W., & Duckworth, P. (2021). Towards better healthcare: What could and should be automated? Technological Forecasting and Social Change, 172, 120967. doi:10.1016/j.techfore.2021.120967.

[22] de Carvalho Junior, M. A., & Bandiera-Paiva, P. (2018). Health Information System Role-Based Access Control Current Security Trends and Challenges. Journal of Healthcare Engineering, 2018, 6510249. doi:10.1155/2018/6510249.

[23] Wang, G.-Y. (2022). Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling. Statistika: Statistics and Economy Journal, 102(4), 443–453. doi:10.54694/stat.2022.18.

[24] Zhang, R., Chen, D., Shang, X., Zhu, X., & Liu, K. (2018). A knowledge-constrained access control model for protecting patient privacy in hospital information systems. IEEE Journal of Biomedical and Health Informatics, 22(3), 904–911. doi:10.1109/JBHI.2017.2696573.

[25] Barad, M. (2019). Linking cyber security improvement actions in healthcare systems to their strategic improvement needs. Procedia Manufacturing, 39, 279–286. doi:10.1016/j.promfg.2020.01.335.

[26] Shukla, A., Katt, B., Nweke, L. O., Yeng, P. K., & Weldehawaryat, G. K. (2022). System security assurance: A systematic literature review. Computer Science Review, 45, 279–286. doi:10.1016/j.cosrev.2022.100496.

[27] Khayrutdinov, M. M., Golik, V. I., Aleksakhin, A. V., Trushina, E. V., Lazareva, N. V., & Aleksakhina, Y. V. (2022). Proposal of an algorithm for choice of a development system for operational and environmental safety in mining. Resources, 11(10), 88. doi:10.3390/resources11100088.

[28] Mitra, A., Soman, B., Gaitonde, R., Singh, G., & Roy, A. (2022). Data science methods to develop decision support systems for real-time monitoring of COVID-19 outbreak. Journal of Human, Earth, and Future, 3(2), 223-236. doi: 10.28991/HEF-2022-03-02-08.

[29] Argaw, S. T., Bempong, N. E., Eshaya-Chauvin, B., & Flahault, A. (2019). The state of research on cyberattacks against hospitals and available best practice recommendations: A scoping review. BMC Medical Informatics and Decision Making, 19(1), 1-11. doi:10.1186/s12911-018-0724-5.

[30] Shaikh, F. A., & Siponen, M. (2023). Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to cybersecurity. Computers & Security, 124, 102974. doi:10.1016/j.cose.2022.102974.

[31] Schmitz, C., Schmid, M., Harborth, D., & Pape, S. (2021). Maturity level assessments of information security controls: An empirical analysis of practitioners assessment capabilities. Computers & Security, 108, 102306. doi:10.1016/j.cose.2021.102306.

[32] Xiang, H., Lu, J., Kosov, M. E., Volkova, M. V., Ponkratov, V. V., Masterov, A. I., Elyakova, I. D., Popkov, S. Yu., Taburov, D. Yu., Lazareva, N. V., Muda, I., Vasiljeva, M. V., & Zekiy, A. O. (2023). Sustainable Development of Employee Lifecycle Management in the Age of Global Challenges: Evidence from China, Russia, and Indonesia. Sustainability, 15(6), 4987. doi:10.3390/su15064987.

[33] Zhao, J., Shao, M., Wang, H., Yu, X., Li, B., & Liu, X. (2022). Cyber threat prediction using dynamic heterogeneous graph learning. Knowledge-Based Systems, 240, 108086. doi:10.1016/j.knosys.2021.108086.

[34] Singh, A., & Chatterjee, K. (2021). Securing smart healthcare system with edge computing. Computers & Security, 108, 102353. doi:10.1016/j.cose.2021.102353.

[35] Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. BMC Medical Informatics and Decision Making, 12(1). doi:10.1186/1472-6947-12-66.

[36] Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation. Advances in Applied Probability, 28(2), 525–539. doi:10.2307/1428070.

[37] Hoshino, N. (2001). Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment. Journal of Official Statistics, 17(4), 499–520.

[38] Zayatz, L. V. (1991). Estimation of the percent of unique population elements on a microdata file using the sample. US Bureau of the Census, Suitland-Silver Hill, United States.

[39] Benedetti, R., & Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. Pre-proceedings of New Techniques and Technologies for Statistics, November, 4-6 November, 1998, Sorrento, Italy.