Case-Based Reasoning in Clinical Processes Using Clinical Data Banks

V.L.Malykh, D.V.Belyshev
Research Center of Medical Informatics
Ailamazyan Program Systems Institute of RAS
Pereslavl-Zalessky, Russia
mvl@interin.ru; belyshev@interin.ru

[1]*Abstract*—**This study describes an approach to building a clinical data bank on the basis of a standardized clinical process model. A clinical data bank is considered as a repository of clinical cases, which is medical knowledge reflecting both up-to-date and retrospective medical data. In view of the case-based nature of the decision making process in medicine, a clinical data bank can be used in clinical processes to make diagnosis and treatment decisions. The study discusses the following issues: a) choosing an approach to modeling a clinical process; b) completeness of the model; c) medical data standardization and normalization; d) plain medical text mining; d) big medical data analysis. The study also looks at metric-based methods of searching for relevant clinical cases in the data bank.**

Keywords— ***diagnostic and treatment process modeling; decision-making system; case-based reasoning; clinical data banks***

I. Introduction

An overview of materials presented at the Fifth World Congress on Medical Informatics discussed medical knowledge formalization [1]. It was proposed to divide medical knowledge into scientific and empirical knowledge. It was noted that models based on scientific medical knowledge are generalized and limited in nature, while empirical knowledge can be insufficiently representative and inadequate for the population in general. Formalized scientific medical knowledge consists of scientific theories and medical experience registered in databases (DB) and knowledge bases (KB). Both components of medical knowledge continued to evolve after the overview was published. Wide-spread automation of diagnosis and treatment processes gave rise to a new class of systems: medical information systems (MIS). MIS strengthened the importance of empirical medical knowledge. MIS DBs have already accumulated empirical knowledge describing and formalizing millions of clinical cases in the form of electronic medical records (EMR) [2]. The scientific component of medical knowledge continues its rapid development, giving rise to new diagnosis and treatment methods as well as new medicines. One of the new trends is personalized medicine, which is tailored to individual risk factors, including genetic ones [3, 4]. On the one hand, medicine is conservative and based on accumulated and proven experience (evidence-based medicine); but on the other hand, it requires that physicians constantly gain new knowledge. Designed to help physicians, decision support systems are facing the problem of outdated knowledge they store; therefore, knowledge must be constantly formalized and updated. This is one of the reasons for slow expansion and practical application of such systems, which has not changed over recent years [1]. It should be noted that the empirical component of medical knowledge includes both advanced and retrospective medical knowledge. It may be generally concluded that support of the empirical component of medical knowledge in MIS requires less effort than formalization and support of the scientific component in MIS. This opens great prospects for the development of medical decision support systems based on the empirical component of medical knowledge. The case-based nature of the management and decision making processes in medicine also favors the use of this approach. If a healthcare professional manages to select (find) a similar (relevant) case of a diagnosis and treatment process (DTP), this case or a set of relevant cases may serve as a basis for forecasting the future course of this DTP, formulating proposals for possible management events or giving comprehensive interpretation of the current state of the process. The main features of case-based reasoning were described in a monograph by N. G. Zagoruiko [5]. “Speaking about “What?” models only, decision-making methods based on certain cases and those based on the general (model) description are equally effective in terms of methodology. Moreover, a model, just like any other generalization tool, disregards some peculiarities of the system’s behavior at each certain point of the decision space. Experience shows that case-based reasoning takes these peculiarities into account, which makes it possible to come to better decisions.” In terms of healthcare, this means that a good physician is always an experienced physician [4]. More and more studies are being published on the use of case-based reasoning in decision support systems [6]. Some researchers state that today we have all the prerequisites for applying case-based reasoning in health informatics [7]. First, now we have IT solutions for working with big data. Second, data sets themselves have accumulated in medical information systems (MIS). The issue of medical decision support can now be solved through application of case-based reasoning. The most important prerequisite for using the case-based approach is the availability of sufficient representative number of clinical cases or, in other words, the availability of a representative clinical data bank.

The study is supported by the Ministry of Education and Science of the Russian Federation (project No. RFMEFI60714X0089) and grant No. 13-07-12012 provided by the Russian Foundation for Basic Research.

978-1-4673-9111-5/15/$31.00 ©2015 IEEE

The key concepts used in the proposed approach to medical decision making will be presented at MEDINFO 2015 [8]. While the required maximum poster size did not allow the authors to describe these concepts in detail, this is done in the study. The description focuses on formalizing the diagnosis and treatment process into a dynamic process model and discusses two such models. In the first model, DTP is modeled by a controlled Markov process. A similar approach is used in some other studies [9]. Authors of another study proposed a Markov model with unique features that make it different from similar ones [10]. First, their study proposes integrating management in the model to record the history of management and "markovize" the process. Second, it proposes applying equivalence relations to medical state characteristics to generalize the description of the state. Both these issues are discussed below. Methodological problems associated with observability of the patient's state prompted the authors to propose another event model of DTP that is not based on the concept of state. Methods of searching for similar relevant cases were proposed for both models.

II. Clinical data bank building issues

A. Assessment of the amount of clinical data

The Program Systems Institute of RAS jointly with Interin Group has spent 20 years researching into health informatics and running automation projects at large Russian healthcare facilities. Interin PROMIS, a medical information system, has accumulated data from about 3*10^6 electronic medical records in its DB. The amount of data on some diseases (the number of clinical cases of the disease) has reached 10^3–10^4 cases in some MIS systems. Experience shows that the number of medical state characteristics plus the number of diagnosis and treatment measures taken by physicians usually makes a total of about 10^3–10^4. The amount of country-wide clinical data is estimated at 10^9–10^10 individual clinical cases.

It should be noted that all this data is derived from MIS systems of certain healthcare facilities. This allows us to view the clinical data bank as a distributed system with ample opportunities for parallel data processing. Big medical data has an advantage over big physical data (LHC): all such data is significant.

B. Clinical data formalization and completeness

Modern medical information systems keep electronic medical records and contain information on millions of various clinical cases formalized to a certain degree. The degree of formalization of the clinical data stored in various MIS systems may vary. One of the problems faced by the Russian health informatics is a lack of mandatory standards for EMR keeping and clinical documents. This hinders semantic tagging (coding) of data in clinical documents. Western countries extensively use SNOMED Clinical Terms for this purpose. Attempts to use these English-language terms in Russia lead to unreasonably complicated management of document models. Some studies argue that potential benefits from translating and localizing SNOMED Clinical Terms into Russian are still unclear [11, 12].

Up-to-date medical information systems model diagnosis and treatment processes as a time sequence of management events (which include ordering diagnostic tests and administering treatment) and monitoring events (which describe the patient's state). Management events are better formalized; hospitals keep statistical and economic records on them, which includes maintaining registers of services rendered, invoicing, planning and centralized control. Medical data associated with monitoring and diagnostic tests is insufficiently formalized. While laboratory test findings are stored in the database of a medical information system in a rather structured way, findings of examination performed using tools and physicians' personal observations are usually stored in the system as plain texts. The inability to perform automated analysis of plain medical texts leads to diagnosis and treatment process models' disregarding the facts contained in these texts and hinders the search for such facts in arrays of clinical data, making the completeness of formalized medical data an issue. One of the methods of solving this problem is plain medical text mining [13, 14]. Russia is now in the beginning of this journey [15, 16].

The main problem about formalizing medical data and building a clinical data bank is how to choose or develop a standardized DTP model common for different diseases.

III. Diagnosis and treatment process models

A. Model involving transitions between states

One of our previous studies describes the possibility of modeling diagnosis and treatment process using controlled stochastic Markov processes [17]. Further development of these ideas helped build a mathematical DTP model [10]. This chapter outlines the basic ideas and provides formalized mathematical proofs from the specified sources.

This model is based on the assumption that DTP is a controlled process. The model sets forth a management notion, u , and a state notion, x . Management means decisions made by the physician and implemented in the future. Management is limited to prescription of various diagnosis and treatment measures by the physician, including diagnostic testing, drug and treatment administration, surgical treatment and manipulation, etc. The physician's choice is based on accumulated medical knowledge about how a certain disease should be treated and on the physician's individual experience. Rather than taking into account all subsets of DTP elements as management options, the physician considers only those subsets that were used in similar situations and proved their clinical effectiveness. Management is explicitly case-based, while its particular goal distinguishes it from the conventional method of management goal setting. A set of possible management values (U) can be considered a given value. However, the process flow function $f(x,u)$ is case-based, i.e. it is known only with regard to cases observed before (x,u).

It was proposed that the model should be built according to the following methodology. First, we shifted from a continuous-time controlled process to a discrete-time process, as patient monitoring and diagnosis and treatment decision making are discrete-time processes. Continuous instrumental patient monitoring is usually aimed at registration of various events, the occurrence of which may be associated with discrete points of time. To denote the sampling step, let us introduce the index variable with its value specified in the superscript. For example, (x^i, u^i) is an event occurring at step No. i of the process. Second, we tried to consider the controlled process memory effect. Management in the current situation (x^i, u^i) will be determined not only by the state (x^i) but also by the whole history of the process and management at earlier steps of DTP $\{i, i-1, i-2, ...\}$. The physician makes decisions based on the life history, disease history, family history, history of allergies and the entire course of the given clinical case. To take account of the DTP memory effect, it was proposed that the data from the above-mentioned histories is included in discrete states of the process, with management integrated. Each element of the diagnosis and treatment process may be associated with a certain integral characteristic of application of this element in DTP. For example, the total dose consumed by the patient will be considered the integral characteristic for a drug and the total radiation dose will be considered the integral characteristic for radiation therapy. Frequency of application of a certain element can also often be considered as an integral characteristic (e.g., the number of electrocardiographic examinations). The formal denotation of the management is as follows: $u^i = \{(c_1^i, u_1), (c_2^i, u_2), ..., (c_m^i, u_m)\}$, where c_j^i is the integral characteristic of the management element u_j at step No. i of the process. Naturally, subtraction (differentiation) and addition (integration) operations are applied to integral characteristics:

$$\Delta u^{i-1} = u^i - u^{i-1} = \{(c_1^i - c_1^{i-1}, u_1), ...(c_m^i - c_m^{i-1}, u_m)\}$$

It should be noted that time was considered as management and was included in the model in the form of its integral characteristic. The proposed approach to registration of the process history allows for supplementing the notion of the current state with the history data and take account of the whole process management history via integral management characteristics.

Third, we proposed generalizing patient's medical state characteristics through determining the equivalence relation for each characteristic, which will break down a set of characteristic values into equivalence classes.

There are certain methodological issues concerning the notion of the patient's medical state [7]. The patient's medical state can be described using many different characteristics. Medical state characteristics are monitored and determined with various frequencies. At an arbitrarily chosen point of time t, it is rather difficult to determine the medical state $x(t)$ precisely. Conventional methods of interpolation and integration of a fast-changing variable are not good enough here. Physicians are often interested in the dynamics of a characteristic rather than in its daily average values. To summarize, various state characteristics are measured at various random points of time with various frequencies, and the measurement findings may become available with a delay.

Another problem is the high-dimensional nature of the space of generalized states, which is composed of the dimensions of management events and the dimensions of the characteristics being monitored. Let us consider additional results of DTP modeling based on real clinical data [7].

TABLE I. SETS OF MODELED DTP.

***ICD*-10 code / disease**	**Number of processes / number of states / number of generalized states / compression ratio / number of state characteristics / number of normalized state characteristics**
J13 / Pneumonia due to Streptococcus pneumoniae	166 / 2,938 / 2,921 / <1% 828 / 128
H26.2 / Complicated cataract	1,255 / 5,778 / 2,308 / 60% 328 / 249
I20.8 / Other forms of angina pectoris	3,069 / 48,909 / 48,513 / <3% 871 / 99
I10 / Essential (primary) hypertension	8,734 / 98,389 / 82,542 / 16% 3,223 / 1,278

Table 1 provides the results of modeling based on real clinical data from DTP sets with regard to four different ICD codes. The sets of processes are sorted by their capacity (from the lowest to the highest).

The table specifies the code and the name of the disease according to the ICD-10. For each of the specified diseases we selected completed clinical cases of in-patient care, where the disease was the primary diagnosis. The table specifies the number of processes (completed cases) selected for the disease. Each of the processes under consideration lasted a whole number of days. The patient's state was registered once a day. Each state was generally described by a set of dimensional characteristics with definite values. A complete range of characteristics registered with regard to a certain set of processes formed a glossary of characteristics typical of the disease. The capacity of four glossaries of characteristics typical of four diseases is presented in the table. Glossaries of characteristics were normalized. The glossary for the J13 disease was standardized by a physician expert. The glossary for the H26.2 disease was normalized by the authors of this article. Preliminary normalization of the glossary for the I20.8 disease involved a statistical approach: all characteristics with a rating below 307 (i.e. registered in less than 10% of processes within this set) were excluded from the glossary for this disease. A similar approach was used for the I10 disease. For the capacity of normalized glossaries of characteristics see Table 1. To assess the compression capability of the process set description it was assumed that any value of each characteristic is grouped into just one equivalence class (maximum compression). Two states were thought to be equivalent if they had one and the same discrete time index of the process step and had the same sets of characteristics describing these two states. For the number of such generic states for each disease see the table. Compression was defined as a percentage by which the number of states decreased from the initial number of states after transition from the initial description of processes to the generalized description. Based on general knowledge, it was assumed that if the modeled set of processes is small, the compression ratio will be insignificant and special features of the process will distinguish it from other dissimilar processes. Such sets of processes include the one related to the J13 disease, which contains just 166 processes: it was compressed by less than 1%. The higher the capacity the modeled set of processes has, the higher the compression ratio is observed: the set of 1,255 processes related to the H26.2 disease was compressed by 60%. The set of 3,069 processes related to the I20.8 disease was compressed by less than 3%, which can be explained by poor normalization of the glossary of characteristics typical of this disease, insufficient capacity of the set of processes, and by the scope of the disease. The I10 disease that includes 8,734 processes was characterized by a relatively low compression of 16%.

The modeling results show that when the generic state space dimension is high (the number of various state characteristics plus the number of various management events), certain process realizations are weakly coupled with each other via shared states. As a matter of fact, the Markov model starts breaking down into numerous separate realizations that are weakly bound to each other.

B. Event model

The above methodological problem can be solved by abandoning the notion of state altogether and excluding it from the model. Let us consider DTP as a sequence of events – observations, their interpretations, diagnosis and treatment decision points – occurring at random discrete points of time. DTP can be formalized as a set of separate sequences of events, with each sequence consisting of homogeneous time-ordered events. The DTP realization index is specified in the superscript.

$\mathrm{T}^{\mathrm{A}} = \{t_1^A, t_2^A, \dots t_{\mathrm{s}}^A \ \}$ is an ordered sequence of points of time when DTP events occurred.

$\mathrm{X}_{\mathrm{j}}^{\mathrm{A}} = \{x_j^A(t_{i1}^A), \dots, x_j^A(t_{in}^A) \ \}, x_j^A \in \mathrm{X}_{\mathrm{j}}, \ j \in \overline{1,l}.$ is a time-ordered sequence of values of characteristic j observed (measured) at the following points of time: $\{ \ t_{i1}^A, t_{i2}^A, \dots, t_{in}^A \ \}, \ t_{ip}^A \in \mathrm{T}^{\mathrm{A}}, \ p \in \overline{1,n}$.

$\mathrm{U}_{\mathrm{k}}^{\mathrm{A}} = \{u_k^A(t_{k1}^A), \dots, u_k^A(t_{km}^A) \ \}, u_k^A \in \mathrm{U}_{\mathrm{k}}, k \in \overline{1,h}.$: a time-ordered sequence of values of the k-component of management at the following points of time: $\{ \ t_{k1}^A, t_{k2}^A, \dots, t_{km}^A \ \}, \ t_{kr}^A \in \mathrm{T}^{\mathrm{A}}, r \in \overline{1,m}.$

Irrespective of which DTP model is chosen, application of the case-based approach to medical decision support systems requires the development of methods for selecting relevant cases from the clinical data bank.

IV. Search for Relevant Clinical Cases in a Clinical Data Bank

A. Metric-based methods of searching for relevant cases

To formalize the notion of similarity between two clinical processes (A and B) or two states of these processes, it is proposed that particular functions representing similarity measures are introduced to all management components and all state characteristics. To simplify the formalization, let us omit explicit mentioning of management, state and time (x, u, t) and create an extended space of states (X) based on these three notions, where the state will be represented as n number of various characteristics (data objects): $x = (x_1, x_2, \dots, x_n)$. Let us define similarity measures as functions for all characteristics of the extended space of states:

$$\mathrm{d}_{\mathrm{j}} : \mathrm{X}_{\mathrm{j}} \times \mathrm{X}_{\mathrm{j}} \to [0,1], \ j \in \overline{1,n} \ ,$$

where X_{j} is a set of values of characteristic j, and

$$\mathrm{d}_{\mathrm{j}}(x_j^A, x_j^B) \in [0,1], \ \mathrm{d}_{\mathrm{j}}(x_j^A, x_j^B) = \mathrm{d}_{\mathrm{j}}(x_j^B, x_j^A),$$
$$\forall x_j \in \mathrm{X}_{\mathrm{j}}, \ \mathrm{d}_{\mathrm{j}}(x_j, x_j) = 0, \ x_j^A, \ x_j^B \in \mathrm{X}_{\mathrm{j}}, \ j \in \overline{1,n}.$$

If possible, the following property should be satisfied:

$$\forall x_j^A, x_j^B, x_j^C \in \mathrm{X}_{\mathrm{j}} \ \mathrm{d}_{\mathrm{j}}(x_j^A, x_j^B) \le \mathrm{d}_{\mathrm{j}}(x_j^A, x_j^C) +$$
$$+ \mathrm{d}_{\mathrm{j}}(x_j^B, x_j^C), \ j \in \overline{1,\mathrm{n}}.$$

It is rather easy to build a metric for a primitive data type (e.g., numerical characteristics); however, when it comes to complex data objects reflecting DTP, this can be very challenging. That is why we do not insist on satisfying this property.

Let us assume that particular similarity measures across all state characteristics and management components are determined. Now let us define the metric in the extended space of states for two states $(x^A, \ x^B)$ as the weighted sum of particular similarity measures.

$$\rho(A,B) = 1 - \frac{1}{m} \sum_{i=1}^{n} \mathrm{d}_i \, (x_i^A, x_i^B) \times c_i,$$

where $\vec{c} = (c_1, c_2, \dots, c_n)$ is a weight vector of characteristics with non-negative components, and $m = \sum_{i=1}^{n} c_i$ is a normalizing constant. Introduction of weight is very important, as it helps take into account the relative information capacity of certain characteristics. Essentially, this allows physicians to build custom metrics while searching for relevant cases in a clinical data bank. Physicians suggested a very good example themselves: “Search the clinical data bank to find all cases of pneumonia in patients who had one kidney removed." It is clear that in this example we are interested in two clinical case characteristics only; all the other characteristics can be omitted by assigning a zero value to their weight in the metric.

B. *Methods of searching for relevant cases*

Let us assume that there is a universal context (K) that describes all clinical cases stored in a clinical data bank. The following types of information are proposed as obvious examples of context variables: 1) the ICD-10 code of the primary diagnosis; 2) identification code of the healthcare facility where the case was treated; 3) patient identification number; 4) physician identification number; 5) duration of the case; 6) outcome of the case; 7) starting date of DTP related to the case; etc.

When searching for relevant cases, the query should be specified through context variables, which helps immediately exclude all cases that do not satisfy these conditions. After the context filtering, the selected cases are analyzed by searching the extended space of states using metric-based methods.

C. *Methods of searching for relevant cases for the model involving transitions between states*

A typical task will be to find clinical cases relevant to a current state (e.g., an incomplete DTP with confirmed diagnosis). In this example, the primary disease is known, which makes it possible to prepare a template with weights assigned to characteristics and choose an appropriate metric in advance. We will look at the processes that are potential relevant cases (the processes that have been selected by context filtering) to select only those states that are relevant with regard to a certain day of stay in the hospital, one state for each process. Then we will have to calculate the metric for pairs of the given current state and the state from the selected case, and rank the selected states according to similarity to the current state. The identified similar states and corresponding DTP will help physicians take diagnosis and treatment decisions and forecast the outcome and length of in-patient care.

To assess the amount of computations, let us use the data from Table 1. For example, a set of 8,734 completed DTP, which included 98,389 states, were modeled for the I10 disease. One case lasted an average of 98,389/8,734 ≈ 11 days of in-patient care. Ranking this set of processes with regards to the given state will require an average of 8,734 metric computations. A prototype of a clinical database based on Oracle DBMS operating on one server (Intel Core 2 Duo Processor E6600, 2.40 GHz, 8 GB RAM, 820 MB DB RAM) performs 1,000 metric computations within 0.047 seconds. Such performance will allow for processing a set of 20,000 cases in less than 1 second.

Pre-determined metrics for certain diseases make it possible to use cluster analysis methods to improve the efficiency of search for relevant cases [5]. The states selected for a specific disease by the in-patient care day are clustered according the given metric. There are various approaches to the search for relevant cases in the metric space. In particular, methods based on small-world graphs seem to be very efficient in terms of speed [18]. This approach can ensure high efficiency when metrics are pre-determined; however, it cannot be applied when a custom metric is used.

D. *Methods of searching for relevant cases used for the event model*

The notion of state is not explicitly defined in the event model. Therefore, the methodology must be based on other notions. The above-mentioned functions can help assess similarity between pairs of values of a certain characteristic of the extended state; or, in terms of the event model, between pairs of homogeneous events. To assess similarity between two clinical processes (A and B), it is necessary to determine associations between the pairs of homogeneous events of processes A and B, compute the metrics between the associated pairs and "convolute" the resulting values according to the number of associated pairs of homogeneous events and weight of events (weight of the extended state characteristics). Since all the events are time-related, it is enough to determine associations between points of time of processes A and B and translate these associations to events, assuming that the pair of homogeneous events is associated once the points of time of these events are associated with each other. For example, let us determine associations between two time sequences, $\mathrm{T}^{\mathrm{A}} = \{t_1^A, t_2^A, \dots t_{\mathrm{s}}^A \}$ and $\mathrm{T}^{\mathrm{B}} = \{t_1^B, t_2^B, \dots t_{\mathrm{w}}^B \}$. Several approaches to this task can be used.

The first approach involves even pacing of processes A and B. Two points of time (t_{i}^A and t_{j}^B) are associated with each other if $](t_i^A - t_1^A)/\Delta h[=](t_j^B - t_1^B)/\Delta h[$ where $]\,\mathrm{n}\,[$ is the integer part of n, and Δh is the pacing of the process.

The second approach is based on the hypothesis about similarity of processes A and B. Two points of time t_{i}^A and t_{j}^B are associated with each other if t_{i}^A is the closest point of time to ${}^*t_{\mathrm{j}}^B$ in the sequence $\mathrm{T}^{\mathrm{A}} = \{t_1^A, t_2^A, \dots t_{\mathrm{s}}^A \}$, where

$${}^*t_{\mathrm{j}}^B = \frac{t_{\mathrm{j}}^B - t_1^B}{t_{\mathrm{w}}^B - t_1^B}(t_s^A - t_1^A) + t_1^A, i \in \overline{1,w}.$$

As a rule, associating points of time from two sequences may lead to the sequences' breaking down into subsequences, giving rise to association between all elements of subsequence $\mathrm{T}_{\mathrm{i}}^{\mathrm{A}} = \{t_{\mathrm{i}}^A, t_{i+1}^A, \dots t_{\mathrm{i+k}}^A \}$ of process A and all elements of subsequence $\mathrm{T}_{\mathrm{j}}^{\mathrm{B}} = \{t_{\mathrm{j}}^B, t_{j+1}^B, \dots t_{\mathrm{j}+m}^B \}$ of process B. Now let us extract corresponding subsequences of values for any selected characteristic X_{j} in processes A and B at points of time $\mathrm{T_i^A}$ and $\mathrm{T_j^B}$. After leaving out the unrepresentative case where characteristic $\mathrm{X_j}$ was not monitored at the specified points of time in processes A and B, we will have two subsequences of values of this characteristic, $\{x_{\mathrm{i}1}^A, x_{i2}^A, \dots x_{\mathrm{ia}}^A\}$ and $\{x_{\mathrm{i}1}^B, x_{i2}^B, \dots x_{\mathrm{ib}}^B\}$, and at least one of these subsequences will be not empty. Now let us supplement the subsequence having fewer elements with conditional indefinite values $x(?)$ in the right-hand side so that the subsequences have equal number of members. $x(?)$ means that the characteristic has an unknown value. Now let us calculate the metrics between pairs of elements of these subsequences (between the first, the second, etc.), sum up the values and divide by the number of pairs – the result will be the distance between these subsequences. After calculating the distance between other pairs of subsequences for the given characteristic in the same way, let us sum up the distances and normalize them by the number of pairs of subsequences. The resulting value with the corresponding weight will be used in the general assessment of distance between processes A and B.

Now we should define the particular similarity measure function $\mathrm{d_j : X_j \times X_j} \rightarrow [0,1]$, $j \in \overline{1,n}$ for the indefinite value $x(?)$. The statistical approach, which is generally used in the statistical decision theory to take uncertainty into account, can help us do this. Let us assume that we know the value distribution function for characteristic $\mathrm{X_j}$ with regard to the given selected set of clinical cases. In this case, the value of the function $\mathrm{d_j}(x_j, x(?)), x_j \in X_j$ can be defined as the expected value of this function provided that $x(?)$ is characterized by the given distribution.

V. Conclusion

The study looks at the case-based approach to medical decision support. To use case-based reasoning in practice, a rather wide representative array of clinical data should be accumulated. The study considers the possibility of building a representative clinical data bank and assesses the required amount of data. It also notes that clinical data standardization, harmonization and completeness issues should be tackled and advanced plain medical text mining techniques should be used. It analyzes two abstract models of the diagnosis and treatment process and proposes methodologies for using these models to search for relevant clinical cases. The study looks at metric-based methods of search, including the possibility of customizing metrics. Besides, it provides the results of computational modeling of diagnosis and treatment processes based on real data.

References

[1] S.N. Butko, V.K. Olshanskiy, New medical support systems abroad, Automation and Remote Control 6 (1990) pp. 3-19.

[2] Y.I. Guliev, The main aspects of the medical information systems development (in Russian), Vrach i informacionnye tehnologii [Physicians and IT] 5 (2014) pp. 10-19.

[3] V. S. Baranov, Genetic passport and the medicine of the future (in Russian), Himiya i zhizn [Chemistry and life] 1 (2011) pp. 6-11.

[4] Standard and personalized medicine in the diagnosis and treatment of patients (in Russian), Voprosyi ekspertizyi kachestva meditsinskoy pomoschi [The issues of examination of quality of medical care] 8 (2014) pp. 23-27.

[5] N.G. Zagorujko, Applied methods of data analysis and knowledge (in Russian) (Novosibirsk: Institut Matematiki im. S.L.Soboleva SO RAN, 1999) [Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, 1999].

[6] P. Varshavsky, R. Alekhin, The method of finding solutions in intellectual decision support systems with case-based reasoning (in Russian), International Journal “Information Models and Analyses”, 4 (2013) pp. 385-392.

[7] V.L. Malykh, Y.I. Guliev, A.V. Eremin, S.V. Rudeckij, Management and decision making in clinical processes (in Russian), XII all-Russian conference on problems of management of VSPU-2014. Moscow: Conference proceedings: 6518.pdf; pp. 6518-6528. http://vspu2014.ipu.ru/node/8581

[8] V. Malykh, Y. Guliev, Precedent approach to decision making in clinical processes, Medinfo’15, 19-23 August 2015, San Paulo, Brazil.

[9] C. Bennett, K. Hauser, Artificial Intelligence Framework for Simulating Clinical Decision-Making: A Markov Decision Process Approach, Artificial Intelligence in Medicine, 1 (2013) pp. 9–19.

[10] V.L. Malykh, Y.I. Guliev, Mathematical model of clinical processes implemented in the class of stohastic Markov control processes with memory (in Russian), Informacionnye tehnologii i vychislitel'nye sistemy [Information technologies and computing systems], 2 (2014) pp. 62-72.

[11] E.S. Pashkina, About systematized nomenclature the medical term SNOMED CT (Problems of completeness, audit, compare, compliance of ontological standards) (in Russian), Vrach i informacionnye tehnologii [Physicians and IT], 2 (2013) pp. 71-78.

[12] T.V. Zarubina, E.S. Pashkina, Prospects for the use of systematized nomenclature of medical terms (SNOMED CT) in Russia (in Russian), Vrach i informacionnye tehnologii [Physicians and IT], 4 (2012) pp. 6-14.

[13] Louhi 2015. The Sixth International Workshop on Health Text Mining and Information Analysis. https://louhi2015.limsi.fr/

[14] Text mining and information analysis of health documents, Artificial Intelligence in Medicine 61 (2014) pp. 127–130.

[15] V.L. Malykh, Y.I. Guliev, A.N. Kalinin, A.V. Kolupaev, S.G. Jurchenko, The possibility of using a speech interface and automatic text processing systems in MIS (in Russian), Vrach i informacionnye tehnologii [Physicians and IT], 5 (2014) pp. 37-47.

[16] A.N. Vinogradov, Y.I. Guliev, E.P. Kurshev, V.L. Malykh, Perspective researches in the field of clinical modeling, control and decision making (in Russian), Vrach i informacionnye tehnologii [Physicians and IT], 5 (2014) pp. 48-59.

[17] V.L. Malykh, Y.I. Guliev, Modeling of medical-diagnostic process in the class of controlled stochastic processes with memory (in Russian), Vrach i informacionnye tehnologii [Physicians and IT], 2 (2013) pp. 6-15.

[18] Y. Malkov, A. Ponomarenko, A. Logvinov, V. Krylov, Approximate nearest neighbor algorithm based on navigable small world graphs, Information Systems, 45 (2014) pp. 61-68.