DEVELOPMENT OF INTELLECTUAL MODELS AND METHODS OF EXPERT SYSTEMS OF CLINICAL MEDICINE

Oleksandr Kuzomin, Oleksii Vasilenko, Tetiana Tolmachova

***Abstract**: The study examines the processes of modeling clinical medicine, the development of analytical methods, the synthesis of biomedical data and knowledge, as well as the creation of expert systems of clinical medicine. We propose ways of developing models and methods for analyzing biomedical Big Data and developing expert-analytical systems of clinical medicine.*

***Keywords**: Open EHR, Web Mining, OLAP, Data Mining, diagnostic criteria, quality data model, natural language processing, cTAKES, conditional random fields.*

***ITHEA Keywords**: J.3 Life and Medical Sciences, D.2.11 Software Architectures, H.3.3 Information Search and Retrieval*

The problem of clinical diagnosis and statement of the research task

The main issue that arises before a doctor when he meets a patient is the diagnosis - a brief medical report on the nature of the disease and the patient's condition.

In the literature [Yager and Mcintyre, 2013, Organization WH, 2011, CMS Quality Data Model, 2015, EHR Incentive Programs, 2017] there is no unified conception of the diagnostic process, its logical structure often describes only the external formal aspects of clinical thinking and logic (Table 1). Textbooks and manuals do not always have a strict system of exposition of the semiotics of diseases and factual material, often there are no clear indications of which symptoms are significant and which are secondary, to what extent these symptoms are permanent and specific for a particular disease.

There are many varieties of clinical diagnoses, they can be formed by the method of construction, by the time of detection, by the degree of validity (Figure 1). In practical medicine, only three types of diagnosis are most often used: preliminary, basic (clinical) and final, which reflect certain stages of diagnosis.

Table 1. Problems of Clinical Medicine

#	The problem of clinical medicine	Direction of research in work
1	Extraction of data and knowledge from the Internet, social networks and medical practices necessary for analytical and algorithmic solutions to the problems of clinical diagnosis	1. Use the electronic health record (EHR) for data normalization. 2. Combining the best results from the clinical archetypes of the Open EHR, guides and ontologies. 3. Using Web Mining: intellectual analysis of data on the Internet. 4. Logical programming to identify knowledge.
2	Choice of models and methods for qualitative, reliable and timely analysis and algorithmization of clinical diagnostics	1. Big Data (using accurate and reliable large samples, IBM Watson Analytic). 2.OLAP 3.Data Mining
3	Multi-criteria analysis and optimization under uncertainty	The use of the Bayesian network of trust
4	Selection of models and methods for processing biomedical Big Data	Hadoop Technologies: 1. Map Only 2. Classic MapReduce 3. Iterative Map Reducse or Map-Collective
5	Ensuring reliable storage of biomedical Big Data	Backup Modification

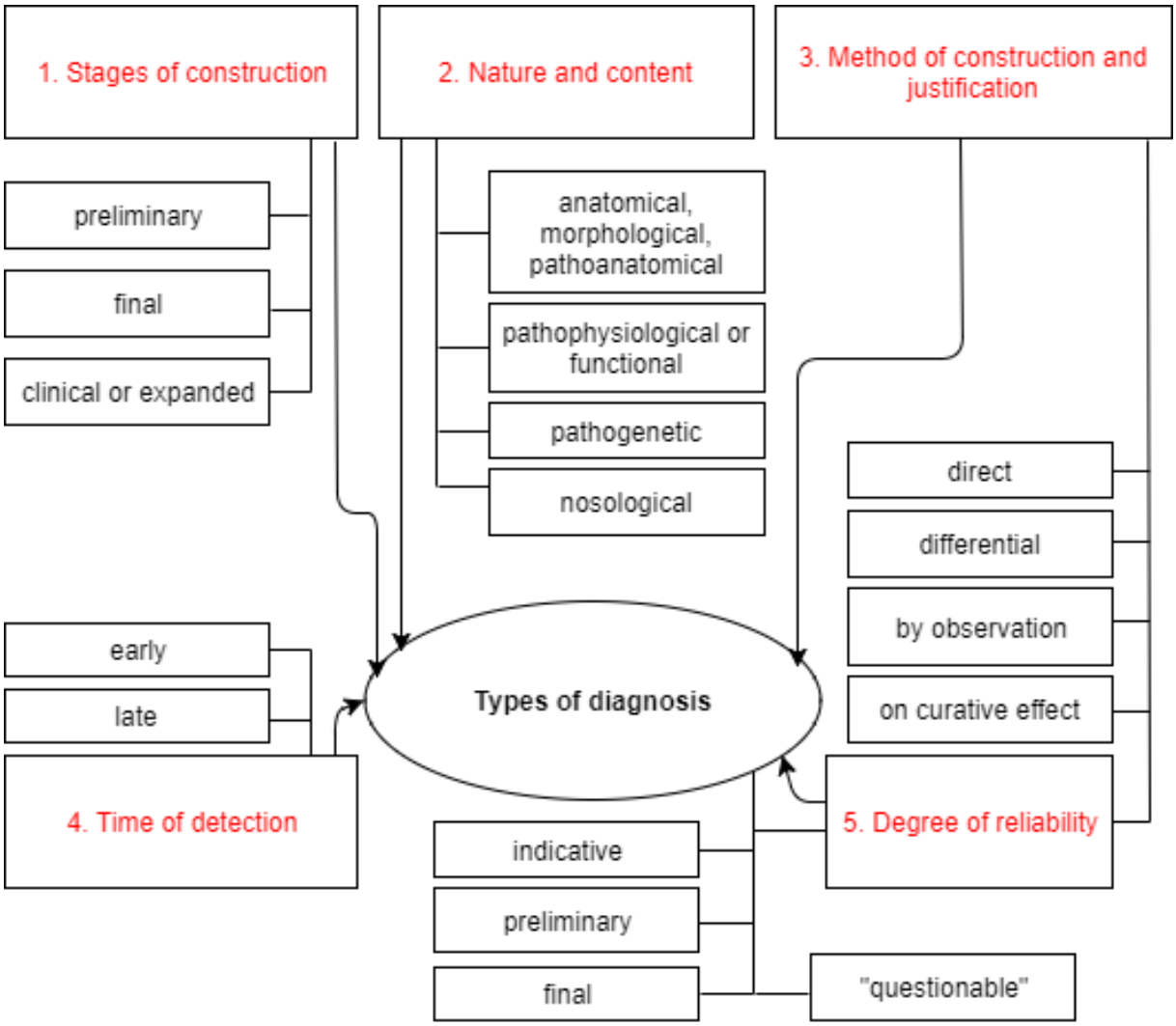

Figure 1. Varieties of medical diagnosis

Logical activity of the doctor-diagnostician is carried out in such forms as concept, judgment, inference, induction and deduction, analysis and synthesis, the creation of ideas and hypotheses. Most often in the diagnosis used inductive method, the method of analogy. If the doctor moves from simple to complex with the use of an inductive method of diagnosis, in the case of analogy, he seeks to learn and recall what he already knows that he has met before, comparing the similarities and differences of symptoms in the observed patients with the symptoms of known diseases.

From the above, we can draw the following conclusion about the problem of diagnosis by traditional methods: the uncertainty of the initial data and the methods of formulating the diagnosis depend to a

large extent on the ontological experience of the doctor, his ability to diagnose using an inductive method, the analogy method. If the doctor moves from simple to complex with the use of an inductive method of diagnosis, in the case of analogy, he seeks to learn and recall what he already knows that he has met before, comparing the similarities and differences of symptoms in the observed patients with the symptoms of known diseases.

The next problem of diagnosis is the ambiguous conformity of symptoms and disease to humans. In addition, diagnosis is often difficult, because many signs and symptoms are nonspecific. Diagnosis is often determined by related symptoms. Often, to model the patient's condition is used Bayesian models. For example, redness of the skin (erythema) in itself is a sign of many disorders and thus does not tell the health care professional what is wrong.

Currently, the most important conclusion about the possibilities of choosing the direction for solving problems in practical medical diagnostics is associated with the use of ontological, precedent Big Data, which are used for their processing with the help of Data Manning analysis, prediction and diagnosis of diseases. As a result, in addition to qualitative diagnostics, it is possible to develop machine learning and diagnostic algorithms.

Ontological, case-law and knowledge databases Big Data analyze textual positive and negative descriptions from medical clinical experience, the Internet and social networks.

Currently, the natural language processing tool (NLP), known as cTAKES [Yager and Mcintyre, 2013, Organization WH, 2011, CMS Quality Data Model, 2015], is used to detect sentences and annotate events in diagnostic criteria. cTAKES consists of several components. Each of them has unique qualities and capabilities. Each component includes at least one analysis engine (annotator), some of which include more. You will want to evaluate the utility of each component for you. UIMA provides a toolkit for selecting which annotators are used together, and how annotators are executed. Each section here covers one component.

cTAKES provides two versions of the source cTAKES pipeline that detects named objects and assigns them attributes:

- For processing text notes: cTAKESdesc / cdpdesc / analysis_engine / AggregatePlaintextProcessor.xml;
- For processing formatted notes in the field of clinical documentation (CDA): cTAKESdesc / cdpdesc / analysis_engine / AggregateCdaProcessor.xml.

Both versions use the same set of components, except that the document preprocessor is not used for plain text.

Suggested solutions

The object of the study are the processes of modeling clinical medicine and the development of methods for analysis and synthesis of biomedical data and knowledge and the creation of expert systems of clinical medicine.

The subject of the study are models and methods of analysis of biomedical Big Data and the development of expert systems of clinical medicine.

The 11th revision of the International Classification of Diseases (ICD-11) was officially published by the World Health Organization (WHO) in March 2007 [EHR Incentive Programs, 2017]. Within the WHO Thematic Advisory Group on Health Informatics and Modeling, a three-level content model was proposed and discussed [Organization WH, 2011]. The purpose of the ICD-11 content model is to present the knowledge underlying the ICD object definitions. Starting in May 2012, the beta phase of the revision of ICD-11 intends to accept public input through a distributed authoring model. One of the main uses is the creation of textual definitions for each category of ICDs. Textual definitions are written down by WHO as follows: "Each ICD concept will be accompanied by a written definition of its descriptive characteristics. This full text definition allows human users to understand the meaning of the concept for classification, translation and other reasons".

The term "diagnostic criteria" refers to a specific sequence of signs, symptoms and test results that clinicians use to determine the correct diagnosis [Kuzomin and Vasylenko, 2014]. This is one of the most valuable sources of knowledge that can be used to support the adoption of clinical decisions and improve patient care [Kuzomin and Vasylenko, 2017]. However, existing diagnostic criteria are scattered across media, such as medical textbooks, literature and clinical practice guidelines, and they are usually described in unstructured free text without a single standard. This situation prevents the effective use of diagnostic criteria to support modern clinical decision-making, which requires an integrated system with interoperable and computable processes.

One of the solutions to better support the adoption of clinical decisions is the computerization of these diagnostic criteria; however, specialists and clinicians are expensive and time consuming to perform all tasks manually. For this purpose, on the one hand, natural language processing technology (NLP) can be used to automatically or semi-automatically convert diagnostic criteria into computable format. On the other hand, the data model for representing diagnostic criteria is equally important for its computerized implementation. Such data model will allow to display diagnostic criteria in a structured,

standard and coded structure to support many clinical applications in a scalable manner. To explore the adaptability of the model to diagnostic criteria, [Kuzomin and Vasylenko, 2014, Kuzomin and Vasylenko, 2010] through a data-based approach in which manually analyzed the distribution and coverage of data elements extracted from a set of diagnostic criteria in QDM. The results showed that the use of QDM is possible when building an information model based on standards for the presentation of computable diagnostic criteria [Kuzomin and Vasylenko, 2010].

The purpose of the study is to develop and evaluate automated methods for converting text clinical diagnostic criteria into a structured format using QDM.

It is proposed to conduct research on the analytical and expert system using biomedical Big Data, obtained as a result of searching the Internet, social networks, literature and the results of practical clinical medicine. Given that the main problem in the formation of biomedical Big Data is the processing of large text documents to facilitate the computerization and standardization of diagnostic criteria, very known clinical tools of NLP and original methods and models are used in the initial stages of the statistical analysis of signs and symptoms of diseases, based on the so-called the concept of microsituations [Kuzomin and Vasylenko, 2017] and self-organizing analytical Data Maning algorithms for textual descriptions of diagnostic data. In particular, we use a combination of known methods of clinical text and knowledge analysis (cTAKES) - supported and rule-based methods for extracting an individual diagnostic criterion from full-text clinical diagnostic criteria and proposed in the study of original methods and models of analysis, forecasting and machine learning. We also develop an algorithm for machine learning based on conditional random fields (CRF) for automatic annotation and classification of the attributes of diagnostic events. Finally, we are developing an integrated web-based system that automatically converts text diagnostic criteria into a standard QDM template by implementing algorithms.

The WHO Content Model ICD-11

WHO developed a content model for presenting the knowledge that underlies the ICD object definition [Kuzomin and Vasylenko, 2010, Yager and Mcintyre, 2013]. The content model consists of three layers:

- The base layer;
- Linearization layer;
- Ontological layer.

The main layer is the main product of the revision of ICD-11, which stores the full range of knowledge about all the classification units in the ICD. Each ICD object can be seen from different dimensions. The content model represents each of these dimensions as a parameter. Currently, 13 basic parameters for the category description in ICD-11 are defined in the content model, as shown in Table. 2.

Table 2. The ICD11 Content Model Main Parameters

#	Parameters
1	ICD Entity Title
2	Classification Properties
3	Textual Definitions
4	Terms
5	Body System/Structure Description
6	Temporal Properties
7	Severity of Subtypes Properties
8	Manifestation Properties
9	Causal Properties
10	Functioning Properties
11	Specific Condition Properties

"Diagnostic criteria" is one of the main parameters for describing the ICD category.

NQF QDM

QDM consists of two modules: a data model module and a logic module. The data model module includes category concepts (eg, medicines), a data type (for example, a "drug controlled"), an attribute (for example, dose, route, strength and duration information) and a set of values containing conceptual codes from one or more terms. The logical module includes logical operators, functions, comparison operators, time operators, subset operators. As mentioned above, HQMF provides a standard format for displaying criteria based on QDM (i.e., instance data) in XML format using a set of templates. [Kuzomin and Vasylenko, 2010, Yager and Mcintyre, 2013, Organization WH, 2011] proposes an assessment of the feasibility of using QDM to represent diagnostic criteria using a data-based approach and the assumption that generic patterns informed by QDM are useful and feasible in constructing a standard-based information model for computable diagnostic criteria. In this study, we used generic templates and selected a set of QDM data types and attributes to develop an initial ontology.

The architecture of the expert-analytical system of clinical diagnosis

Based on the accepted general structure of the system for creating criteria for a clinical diagnosis based on rules is shown. The system architecture contains two main modules: one is a development module that uses a standard information model, and the other is a translation module that uses SWRL. The first architecture module contains an initial ontology that supports the organization of the elements of the diagnostic criteria. The collection of the collection of the manually selected elements of the ICD-11 content model and QDM elements is used, which were informed by an analysis of real diagnostic criteria. The first module also contains a unified web user interface that supports the collection and development of diagnostic criteria by clinicians or subject matter experts on the Internet. The standard QDM model serves as the base layer for translation and reasoning. All collected data items, value sets, and logical expressions of diagnostic criteria are formalized using HQMF templates based on QDM. The second architecture module contains a rule transformation mechanism that converts the diagnostic criteria presented in the QDM / HQMF format into an ontology of the diagnostic criteria for a particular domain and a set of rules using SWRL.

The model of the diagnostic environment at time t:

$$P' \to Mod_{\text{ПС}} \leftrightarrow Opt_{\text{Я}}(P'_1, P'_2, \dots, P'_j, \dots, P'_{13}) \quad (1)$$

where $P' = \{P'_1, P'_2, \dots, P'_{13}\}$ - a set of statistical and analytical procedures for the transformation of information; P'_1 - procedure of multifactor analysis and ranking of parameters of the diagnostic environment; P'_2 - procedure for regression analysis of medical signs; P'_3 - procedure for the variance analysis of medical features; P'_4 - procedure for classifying situations by referring to a confirmed diagnosis and unconfirmed diagnosis; P'_5 - procedure of cluster analysis of symptoms; P'_6 - Data Mining procedure; P'_7 - procedure for compiling a tree for selecting diagnostic solutions; P'_8 - the GMDH procedure; P'_9 - disease pattern recognition procedures; P'_{10} - procedure using fuzzy logic; P'_{11} - procedure for the determination and use of many micro-situations; P'_{12} - procedure for assessing the risk of disease diagnosis; P'_{13} - procedure for synthesizing methods of analysis, forecasting, machine learning.

$$P'_{11} \rightarrow \{Diag'\} \tag{2}$$

where $\{Diag'\}$ - set of medical diagnostic micro-situations.

As a result of the conducted studies based on the results of preliminary analysis of a priori data built a diagnostic model environment for the symptom j and a parameter i:

$$Diag^{j}_{\Pi C_i} = f(X_1^{1,f}, X_2^{2,f}, \dots, X_i^{j,f}, \dots, X_N^{M,f}) \tag{3}$$

where $j = \overline{1, M}$ - number of diagnostic micro-situations, M' - number of symptoms when $M \in M'$, $\{M'_1, M'_2, \dots, M'_f, \dots, M'_F\}$ - set of symptoms, $i = \overline{1, N}$ - number of medical signs of the disease, $\{X_i^{j,f}\}$ - a variety of medical signs of the disease diagnostic environment.

A generalized model of a problematic medical diagnosis $Diag_t$ at a controlled point in time t in:

$$Diag_t = \begin{cases} Diag'_{1,1} = f(X_{1,1}^{M'_1}, X_{2,1}^{M'_1}, \dots, X_{i,1}^{M'_1}, \dots, X_{N,1}^{M'_1}) \\ Diag'_{2,2} = f(X_{1,2}^{M'_2}, X_{2,2}^{M'_2}, \dots, X_{i,2}^{M'_2}, \dots, X_{N,2}^{M'_2}) \\ \dots \\ Diag'_{j,d} = f(X_{1,d}^{M'_f}, X_{2,d}^{M'_f}, \dots, X_{i,d}^{M'_f}, \dots, X_{N,d}^{M'_f}) \\ \dots \\ Diag'_{M,t} = f(X_{1,t}^{M'_F}, X_{2,t}^{M'_F}, \dots, X_{i,t}^{M'_F}, \dots, X_{N,t}^{M'_F}) \end{cases} \quad (4)$$

Development of a standard diagnostic criterion for the initial ontology

The goal here is to integrate existing standard information models relevant to the modeling of diagnostic criteria, through examination and expert editing. As mentioned earlier, we choose the content model ICD-11 and NQF QDM as reference models. The proposed study at this stage is to create an initial ontology of diagnostic criteria (DCUO) by integrating the ICD-11 content model with those QDM elements that are commonly used in diagnostic criteria. The distribution of QDM elements is evaluated using a set of textual diagnostic criteria. The selection of these QDM elements was confirmed by the results of the study, where 10 types of QDM data and 4 QDM attributes were selected and integrated with the ontology scheme based on the ICD-11 content model. In Table 3 provided a list of QDM data types and attributes used for integration.

Table 3. A list of selected QDM datatypes and attributes for developing the upper ontology

QDM Datatypes	QDM Attributes
Laboratory Test, Result	Result
Diagnostic Study, Performed	Method
Diagnostic, Active	Reason
Physical Exam, Performed	Severity

Symptom, Active	
Medication, Active	
Patient Characteristic Birth Date	
Patient Characteristic Race	
Patient Characteristic Sex	
Procedure, Recommended	

We used the Protégé ontology editing environment to manually integrate these two standard information models into the upper ontology of the diagnostic criteria.

The result of the work of the expert system in the form of a list of diagnoses using the ontology of the human phenotype (HPO) is presenting.

Modeling is based on four main stages: 1) a case reduced from HPO, by extracting from the ontology modules related to SARA, 2) an annotation from the free text description scoring scale with ontological modules, 3) development of two clinical archetypes, observation (for normalization content of the scale) and evaluation (for registration of clinical interpretations), 4) identification of information processing units for expressing the system for supporting clinical interpretation

Model approach:

1. Extraction of a shortened version of HPO;
2. Annotations to free text descriptions of elements and SARA estimates;
3. Development of two archetypes (observation and evaluation);
4. Definition of information processing.

Representation of knowledge using the Bayesian network of trust and conditional independence of events is shown on Figure 2.

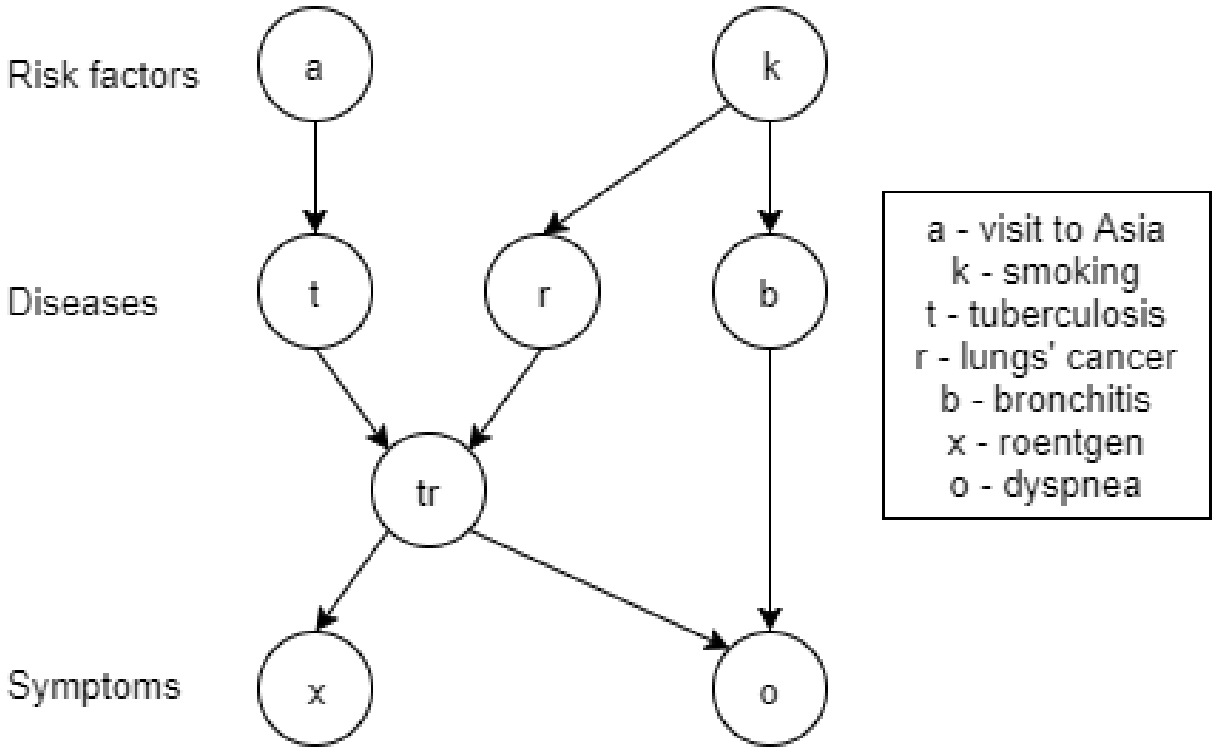

Figure 2. Knowledge representation

Representation of the fragment of the model of the medical DB in the form of BSD. This model corresponds to the following set of medical knowledge:

Dyspnea [o] may be due to tuberculosis [t], lung cancer [r] or bronchitis [b], as well as due to none of the listed diseases or more than one.

A visit to Asia [a] increases the chances of tuberculosis [t].

Smoking [k] is a risk factor for both cancer [r] and bronchitis [b].

X-ray results, determining the darkness in the lungs, do not distinguish between cancer [r] and tuberculosis [t], nor does it determine the presence or absence of dyspnoea [o].

The last fact is represented in the graph as an intermediate variable (event) [tr]. This variable corresponds to the logical function "or" for two parents ([t] and [r]) and it means the presence of either one or two diseases or their absence.

An important concept of the Bayesian network of trust is the conditional independence of the random variables corresponding to the vertices of the graph.

Two variables A and B are conditionally independent at a given third vertex C, if for a known value of C, the value of B does not increase the informativity about the values of A, that is,

$$p(A|B,C) = p(A|C) \quad (5)$$

If there is a fact that the patient is smoking, then we establish our trust regarding cancer and bronchitis. However, our trust regarding tuberculosis does not change. That is, [t] does not conditionally depend on [k] for a given empty set of variables

$$p(t|k) = 0 \quad (6)$$

Receiving a positive result of a patient's X-ray increases our confidence in tuberculosis and cancer, but not about bronchitis. That is, [b] - conditionally does not depend on [x] for a given k:

$$p(b|x,k) = p(b|k) \quad (7)$$

However, if we also knew that the patient had frequent breathing [o], then the x-ray results would also have an effect on our confidence in bronchitis. That is, [b] conditionally depends on [x] for given o and k. Thus, the inference in BDB means the calculation of conditional probabilities for some variables in the presence of information (evidence) about other symptoms.

The construction of standard and computable clinical diagnostic criteria is an important but challenging area of research in the community of clinical informatics. The Quality Data Model (QDM) is becoming a promising information model for standardizing clinical diagnostic criteria.

Conclusion

The solution of the set tasks allowed to obtain such results:

- Develop a method for distributing clinical medical situations to dangerous and safe classes that determines the most important factors of influence on the clinical state of the patient, which provides an opportunity to propose a technique for predicting the patient's condition using the proximity measure of micro-situations (with the greatest influence of parameters);

- A reliable backup method for storage of biomedical Big Data has been developed, which in practice has ensured high reliability in comparison with existing methods;
- A further development of the method for constructing the structure of expert systems for clinical medicine, which is distinguished by the repeated use of the ontology of successful outcomes, makes it possible to improve the reliability and speed of processing input data, as well as the effectiveness of decision-making;
- The situational model of clinical medicine has been improved for the analysis of the patient's condition, which, unlike existing approaches, uses situational presentation of the crisis situation on the basis of the triple "doctor-control action or solutions for re-use of the ontology-patient", which allows forecasting dangerous and safe situations for patient faster and with greater accuracy than existing models.

Bibliography

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Data loss minimization in situation's centrums databases. Chairman IDRC Davos 2014, Global Risk Forum GRF Davos, Davos, Switzerland. pp. 153-154

[Kuzomin and Vasylenko, 2017] Kuzomin, O.Ya., Vasylenko, O., Methods and models for building a distributed mobile emergency monitoring system. 17th International Multidisciplinary Scientific Geoconference SGEM 2017, Conference Proceedings, Informatics Geoinformatics, Vol.17. 2017. pp. 433 – 440.

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Obespechenie bezopasnosti ispolzovaniia baz dannyh v usloviiah chrezvychainyh situazii. International Journal "Information Technologies Knowledge", Vol. 8, Num. 2. 2014. pp. 173-187.

[Kuzomin and Vasylenko, 2010] Kuzomin, O.Ya., Vasylenko, O., Analiz estestvenno iazykovyh obiektov I predstavlenie znanii. Vostochno-Evropeiskii zhurnal peredovyh technologii, Vol. 6/2(48). 2010.

[Yager and Mcintyre, 2013] Yager, J., Mcintyre, J.S.. DSM-5 clinical and public health committee: challenges and considerations. Am J Psychiatr, 2014, 171(2). pp. 142–144.

[Organization WH, 2011] ICD-11 Alpha Content Model Reference Guide, 11th Revision. Geneva, Switzerland: World Health Organization. 2011.

[CMS Quality Data Model, 2015] CMS. Quality Data Model,Version 4.2. 2015. https://ecqi.healthit.gov/system/files/qdm_4_2_aug_2015.pdf.

[EHR Incentive Programs, 2017] Electronic Health Records (EHR) Incentive Programs. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html.

[ICD-11 Information Models, 2013] ICD-11 Information Models. 2013. http://informatics.mayo.edu/icd11model.

Authors' Information

Prof. Dr.-hab. Oleksandr Kuzomin *– Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine; e-mail: kuzy@daad-alumni.de; tel.: +38(057)7021515*

Major Fields of Scientific Research: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.

Oleksii Vasilenko – *Data Strategy Manager, Australia, Aspirant of Kharkov National University of Radio Electronics; Kharkiv, Ukraine; e-mail: ichbierste@gmail.com ; tel.: +380 63 841 66 23*

Major Fields of Scientific Research: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics

Tetiana Tolmachova *–Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: tetiana.tolmachova@gmail.com*

Major Fields of Scientific Research: Big Data, Data Mining, Data Analyses.